

# Borrowing strength in hierarchical Bayes: convergence of the Dirichlet base measure <sup>1</sup>

XuanLong Nguyen  
xuanlong@umich.edu

Technical report 532  
Department of Statistics  
University of Michigan

January 4, 2013  
(This version: November 2, 2013)

## Abstract

This paper studies posterior concentration behavior of the base probability measure  $G$  of a Dirichlet measure  $\mathcal{D}_{\alpha G}$ , given observations associated with  $m$  Dirichlet processes sampled from  $\mathcal{D}_{\alpha G}$ , as  $m$  and the number of observations  $m \times n$  tend to infinity. The base measure itself is endowed with another Dirichlet prior, a construction known as the hierarchical Dirichlet processes [Teh et al., 2006]. Convergence rates are established in transportation distances (i.e. Wasserstein metrics) under various geometrically sparse conditions on the support of the true base measure. As a consequence of the theory we demonstrate the benefit of “borrowing strength” in the inference of multiple groups of data — a heuristic argument commonly used to motivate hierarchical modeling. In certain settings, the gain in efficiency due to the latent hierarchy can be dramatic, improving from a standard nonparametric rate to a parametric rate of convergence. Tools developed include transportation distances for nonparametric Bayesian hierarchies of random measures, the existence of tests for Dirichlet measures, and geometric properties of the support of Dirichlet measures.

## 1 Introduction

Ferguson’s Dirichlet process is a fundamental building block in nonparametric Bayesian statistics [Ferguson, 1973, Blackwell and MacQueen, 1973, Sethuraman, 1994]. Recent advances in modeling and computation have seen Dirichlet processes routinely built into hierarchical probabilistic structures in innovative ways [Hjort et al., 2010]. A particularly interesting structure that is also the focus of this paper, is the hierarchical Dirichlet processes [Teh et al., 2006, Teh and Jordan, 2010], where the base probability measure of the

---

<sup>1</sup> AMS 2000 subject classification. Primary 62F15, 62G05; secondary 62G20.

Key words and phrases: Dirichlet process, geometry of support, hierarchical Dirichlet processes, base measures, Wasserstein metric, optimal transportation, posterior consistency, rates of convergence

This research was supported in part by NSF grants CCF-1115769 and OCI-1047871.

Dirichlet becomes an object of inference, which is endowed with yet another Dirichlet prior. The hierarchical Dirichlet processes have been successfully applied to clustering of grouped data in a vast number of application domains.<sup>2</sup>

This paper investigates the asymptotic behavior of measure-valued latent variables that arise in the hierarchical Dirichlet processes. The basic question that we address is the convergence of an estimate of the base probability measure (hereafter “base measure”) of a Dirichlet measure, given observations associated with the Dirichlet processes sampled by the Dirichlet. Let  $\Theta$  be a complete separable metric space equipped with the Borel sigma algebra,  $\mathcal{P}(\Theta)$  the space of probability measures on  $\Theta$ , and let  $G \in \mathcal{P}(\Theta)$  and  $\alpha > 0$ . Recall from Ferguson [1973] that a Dirichlet process  $Q$  is a random measure taking value in  $\mathcal{P}(\Theta)$  and distributed by a Dirichlet measure  $\mathcal{D}_{\alpha G}$ , if for any measurable partition  $(B_1, \dots, B_k)$  of  $\Theta$  for some  $k \in \mathbb{N}$ , the random vector  $(Q(B_1), \dots, Q(B_k))$  is distributed according to the  $k$ -dimensional Dirichlet distribution with parameters  $(\alpha G(B_1), \dots, \alpha G(B_k))$ .

**Problems.** Let  $Q_1, \dots, Q_m$  be an iid  $m$ -sample from a Dirichlet measure  $\mathcal{D}_{\alpha G}$ , where  $\alpha > 0$  is given and the base measure  $G = G_0$  is unknown. By a basic property of Dirichlet processes,  $Q_1, \dots, Q_m$  are random measures on  $\Theta$  that are discrete with probability one. They will *not* be assumed to be observed directly. Instead, for each  $i = 1, \dots, m$ , we shall be given an iid  $n$ -sample  $Y_{[n]}^i = (Y_{i1}, \dots, Y_{in})$  from a mixture distribution in which  $Q_i$  serves as a mixing measure. This mixture distribution admits the density function  $p_{Q_i}(x) := Q_i * f(x) = \int f(x|\theta)Q_i(d\theta)$ , where  $f(\cdot|\cdot)$  is a known kernel density function defined with respect to a dominating measure on  $\Theta$ .

To estimate  $G_0$  by taking a Bayesian approach, the base measure  $G$  is endowed with a prior on the space of measures  $\mathcal{P}(\Theta)$ , yielding a hierarchical model specification as follows:

$$G \sim \Pi_G, \quad Q_1, \dots, Q_m | G \stackrel{iid}{\sim} \mathcal{D}_{\alpha G}, \quad (1)$$

$$Y_{i1}, \dots, Y_{in} | Q_i \stackrel{iid}{\sim} Q_i * f, \quad \text{for } i = 1, \dots, m. \quad (2)$$

For the choice of prior  $\Pi_G := \mathcal{D}_{\gamma H}$ , where  $\gamma > 0$  and  $H \in \mathcal{P}(\Theta)$  is non-atomic and known, this construction is called the hierarchical Dirichlet processes model [Teh et al., 2006]. Fast computational methods have been developed to collect samples from the posterior distributions of interest, such as those for the latent  $G$  and  $Q_i$ ’s, given the  $m \times n$  data set  $Y_{[n]}^{[m]} := (Y_{[n]}^1, \dots, Y_{[n]}^m)$ . The first question considered in this paper is the following:

- (I) How fast does the posterior distribution of the base measure  $G$  concentrates toward the true  $G_0$ , as  $n$  and  $m$  tend to infinity?

An appealing aspect well appreciated by modelers and practioners of hierarchical modeling is the notion of “borrowing strength”. Latent variables shared higher up in a conditional independence probability hierarchy provide an infrastructure through which one

---

<sup>2</sup>Google scholar page shows more than 1400 citations of Teh et al. [2006].

may improve the inference of a parameter of interest by borrowing from information on other related data and parameters also included in the model. For the hierarchical Dirichlet processes, the “borrowing” has a concrete meaning: according to the model, the Dirichlet processes  $Q_i$ ’s for all  $i = 1, \dots, m$  share the same set of supporting atoms as that of the base measure  $G$ . It is intuitive that the inference of the supporting atoms of, say,  $Q_1$  for group 1, should benefit from the information given by other groups of data associated with  $Q_2, Q_3$  and so on. To quantify this intuition, we ask the following:

- (II) What is the posterior concentration behavior of a mixture distribution, denoted by  $Q * f$ , as  $Q$  is attached to the Bayesian hierarchy in the same way as the  $Q_i$ ’s, in comparison to a “stand-alone” mixture model  $Q * f$ , where  $Q$  is endowed with an independent prior distribution.

By resolving question (I), we can demonstrate situations in which the Bayesian hierarchy has the effect of translating the posterior concentration behavior of base measure  $G$  to improved posterior concentration behavior of each individual group of data in the setting of question (II). Both questions will be addressed using the tools that we develop with transportation distances [Villani, 2008].

**Related Works.** The only work known to us about the inference of the Dirichlet base measure is by Korwar and Hollander [1973], who show that it is possible to obtain a consistent estimate (in some sense) of a base measure  $G_0$ , given an iid  $n$ -sample from  $m = 1$  Dirichlet process  $Q_1$  distributed by  $\mathcal{D}_{\alpha G_0}$ . This somewhat surprising result is due to two crucial assumptions made in their work: the true base measure  $G_0$  is non-atomic, *and*  $Q_1$  is observed directly. Due to the fact that two Dirichlet measures with different non-atomic base measures are orthogonal, the estimation of non-atomic base measures becomes somewhat trivial if the sampled Dirichlet processes  $Q_i$ ’s are observed directly. Changing at least one of the two assumptions makes the question considerably more difficult, which leads to different answers and requires new proof techniques. In this paper, we study the case  $G_0$  is an atomic measure with either finite or infinite support, and the  $Q_i$ ’s are *not* observed directly. To get a sense of the challenge, consider the simplest case, that the base measure  $G_0$  has a finite number of support points, say  $G_0 = \sum_{i=1}^k \beta_i \delta_{\theta_i}$ , where  $\theta_1, \dots, \theta_k$  are *known*. Having a single observation  $Q_1$  distributed by  $\mathcal{D}_{\alpha G_0}$  is equivalent to having a single draw from a  $k$ -dim Dirichlet distribution with parameter  $(\alpha\beta_1, \dots, \alpha\beta_k)$ . It is clearly impossible to obtain a consistent estimate of  $G_0$  by setting  $m = 1$  (or finite), and  $n \rightarrow \infty$ . In addition, the assumption that  $Q_1, \dots, Q_m$  are *not* observed directly makes the analysis considerably more delicate, due to the fact that we no longer have access to a simple point estimate of the Dirichlet base measure, as allowed in Korwar and Hollander [1973]. We leave open the setting where  $G_0$  is non-atomic *and* the  $Q_i$ ’s are not observed directly. For this setting, the Dirichlet prior in the hierarchical Dirichlet processes may not be a good choice, due to the discreteness of Dirichlet processes. On the other hand, there is no known practical estimation method available for this setting at the moment.

The convergence theory of posterior distributions has received much development in the past decade. Recent references include Barron et al. [1999], Ghosal et al. [2000], Shen and Wasserman [2001], Ghosh and Ramamoorthi [2002], Walker [2004], Ghosal and van der Vaart [2007], Walker et al. [2007]. See Ghosal [2010] for a concise overview. This theory when applied to density estimation problem has become quite mature — the dominant theme is a Hellinger theory of density estimation for observed data. On the other hand, asymptotic behaviors of latent variable models remain poorly understood. When the inference of a latent variable is of primary concern, the Hellinger theory is not adequate, because one has to account for the underlying geometry of the variables of interest. There are some examples of such asymptotic theory that have been developed recently, e.g., for models of random functions [van der Vaart and van Zanten, 2008, Giné and Nickl, 2011], models of mixing measures [Rousseau and Mengersen, 2011, Nguyen, 2013a] and models of random polytopes [Nguyen, 2013b]. In particular, in a prior work the author demonstrated the usefulness of Wasserstein distances in analyzing the convergence of latent mixing measures in mixture models [Nguyen, 2013a]. This viewpoint will be deepened in this work for a canonical class of nonparametric and hierarchical models equipped with a general class of optimal transport distances for hierarchies for random measures.

Latent hierarchies have long been a versatile and highly effective modeling tool for statistical modelers (see, e.g., Berger [1993]). They can also be viewed as a device for frequentist concepts of shrinkage and random effects (see, e.g., Chapter 5 of Lehmann and Casella [1998]). Due to their wide usages, it is of interest to characterize the roles of latent hierarchies and their effects on posterior inference in a rigorous manner. Theoretical work addressing such questions, particularly for nonparametric and hierarchical models, remains rare in the literature.

**Overview of results.** This paper presents several contributions to the Bayesian asymptotics literature for hierarchical and nonparametric models, as illustrated by our study of the hierarchical Dirichlet processes. Primary contributions include: (1) an analysis of convergence for the estimation of the base measure (mean measure) of a Dirichlet measure, addressing a line of inquiry started by the early work of Korwar and Hollander [1973]; (2) a theoretical analysis of the effect of “borrowing of strength” in the latent nonparametric hierarchy of variables, presenting a step forward in the asymptotic treatment of hierarchical models. In addition, as part of the proofs of these two results we develop new tools that help to shed light on the geometry of the support of Dirichlet measures, and the geometry of test sets that discriminate among different Dirichlet measures. As mentioned earlier, our geometric theory is equipped with Wasserstein distances, and a new class of transportation distances that we will introduce.

Recall that for  $r \geq 1$ , the  $L_r$  Wasserstein distance between two probability measures  $G, G' \in \mathcal{P}(\Theta)$  is given as

$$W_r(G, G') = \inf_{\kappa \in \mathcal{T}(G, G')} \left[ \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}. \quad (3)$$

Here,  $\mathcal{T}(G, G')$  is the space all joint distributions on  $\Theta \times \Theta$  whose marginal distributions are  $G$  and  $G'$ . Such a joint distribution  $\kappa$  is also called a coupling between  $G$  and  $G'$  [Villani, 2008].

Our first main result (Theorem 2.1 in Section 2) establishes the following. Suppose that the  $m \times n$  data set  $Y_{[n]}^{[m]} := (Y_{[n]}^1, \dots, Y_{[n]}^m)$  are generated by the model specified by Eqs. (1) and (2), according to  $G = G_0$  for some unknown  $G_0 \in \mathcal{P}(\Theta)$ . As  $n \rightarrow \infty$ , also  $m \rightarrow \infty$  at a specified rate with respect to  $n$ , there is a net of scalars  $\epsilon_{mn} \downarrow 0$  such that the posterior probability

$$\Pi_G \left( W_1(G, G_0) \geq \epsilon_{mn} \middle| Y_{[n]}^{[m]} \right) \longrightarrow 0 \quad (4)$$

in  $P_{Y_{[n]}|G_0}^m$ -probability. Here,  $P_{Y_{[n]}|G_0}^m$  denotes the true probability measure that generates the data set. The concentration rate  $\epsilon_{mn}$  is made of two quantities  $\epsilon_{mn} = A_1 + A_2$ . In particular,  $A_1 \asymp [n^d \log(mn)/m]^{1/(d+2)}$ , which tends to zero as long as  $m \gg n^d \log n$ . Quantity  $A_2 \rightarrow 0$  as  $n \rightarrow \infty$ , and can be defined as a function of the *demixing* rate  $\delta_n$  of a deconvolution problem (cf. Carroll and Hall [1988], Zhang [1990], Fan [1991]). To be clear,  $\delta_n$  is the rate of convergence — in  $W_2$  in our case — for estimating the mixing measure given an iid  $n$ -sample of a mixture density with kernel  $f$ . The nature of the dependence of  $A_2$  on  $\delta_n$  is interesting, as it hinges on the geometry of the support of the true base measure  $G_0$ . We can establish a sequence of gradually deteriorating rates as the support of  $G_0$  becomes less sparse:

- (i) if  $G_0$  has a finite and known number of support points on a bounded subset of  $\mathbb{R}^d$ , then  $A_2 \asymp \delta_n^{\alpha^*}$ . In fact, one obtain the overall parametric rate of convergence under some conditions and constant  $\alpha^* = \inf_{\theta \in \text{spt } G_0} \alpha G(\{\theta\})$ , that  $\epsilon_{mn} \asymp [\log(mn)/m]^{1/2} + [(\log n)^{1/2}/n^{1/4}]^{\alpha^*}$ .
- (ii) if  $G_0$  has a finite and unknown number of support points on a bounded subset of  $\mathbb{R}^d$ , then  $A_2 \asymp \delta_n^{\alpha^*/(\alpha^*+1)}$ .
- (iii) if  $G_0$  has an infinite number of geometrically sparse support points on a bounded subset of  $\mathbb{R}^d$ , then  $A_2 \asymp \exp -[\log(1/\delta_n)]^{1/(1 \vee \gamma_0 + \gamma_1)}$  for “supersparse” measures, or  $A_2 \asymp [\log(1/\delta_n)]^{-1/(\gamma_0 + \gamma_1)}$  for “ordinary sparse” measures.

The notion of ordinary and supersparse measures mentioned in (iii) will be defined in Section 2. At a high level they refer to probability measures that have geometrically sparse support on  $\Theta$ , where the sparseness is characterized in terms of parameters  $\gamma_0$  and  $\gamma_1$ , which are respectively analogous to the Hausdorff dimension and the packing dimension that arise in fractal geometry [Falconer, 1985, Garcia et al., 2007].

A notable feature about this result is the interaction between quantities  $m$  and  $n$ . They play asymmetric roles in the model hierarchy:  $m$  is the number of groups of data, and  $n$  is the sample size for each group. The appearance of  $n$  in  $A_1$  suggests that if a group of  $n$  represents a “data point”  $Q_i$  (in an iid  $m$ -sample of  $Q_i$ ’s), then  $n$  may be viewed as a sort of dimensionality for each data point. One can draw a parallel between the  $(m, n)$  asymptotic

setting for a hierarchical model and that of  $m$ -sample,  $n$ -dimension in high-dimensional inference [Buhlmann and van de Geer, 2011]. This analogy is not entirely accurate, of course, because the  $Q_i$ 's are not observed directly, have potentially infinite dimensions, and for which better estimates may be obtained as  $n$  increases. This is a distinct feature that separates hierarchical models from other standard models in the literature.

Our second main result establishes the effect of “borrowing strength” of hierarchical modeling. Suppose that an iid  $\tilde{n}$ -sample  $Y_{[\tilde{n}]}^0$  drawn from a mixture model  $Q_0 * f$  is available, where  $Q_0 = Q_0^* \in \mathcal{P}(\Theta)$  is unknown:

$$Y_{[\tilde{n}]}^0 | Q_0 \stackrel{iid}{\sim} Q_0 * f. \quad (5)$$

In a stand-alone setting  $Q_0$  is endowed with a Dirichlet prior:  $Q_0 \sim \mathcal{D}_{\alpha_0 H_0}$  for some known  $\alpha_0 > 0$  and non-atomic base measure  $H_0 \in \mathcal{P}(\Theta)$ . Under mild conditions on the Dirichlet process mixture, it can be shown that in Hellinger metric, the posterior probability

$$\Pi_Q \left( h(Q_0 * f, Q_0^* * f) \geq C(\log \tilde{n}/\tilde{n})^{\frac{1}{d+2}} \middle| Y_{[\tilde{n}]}^0 \right) \longrightarrow 0 \quad (6)$$

in  $P_{Y_{[\tilde{n}]}^0 | Q_0^*}$ -probability for some constant  $C > 0$ . Alternatively, suppose that  $Q_0$  is attached to the hierarchical Dirichlet process in the same way as the  $Q_1, \dots, Q_m$ , i.e.:

$$G \sim \mathcal{D}_{\gamma H}, \quad Q_0, Q_1, \dots, Q_m | G \stackrel{iid}{\sim} \mathcal{D}_{\alpha G}. \quad (7)$$

Implicit in this specification, due to a standard property of the Dirichlet, is the assumption that  $Q_0$  shares the same set of supporting atoms as  $Q_1, \dots, Q_m$ , as they share with the (latent) discrete base measure  $G$ .

Theorem 2.2 in Section 2 establishes the posterior concentration rate  $\delta_{m,n,\tilde{n}}$  for the mixture density  $Q_0 * f$ , under the hierarchical model given by Eq. (7), as  $\tilde{n} \rightarrow \infty$  and  $m, n \rightarrow \infty$  at suitable rates. Specifically, suppose that the true base measure  $G_0$  has a finite number of support points, if  $m$  and  $n$  grow sufficiently fast relatively to  $\tilde{n}$  so that the base measure  $G$  converges to  $G_0$  at a sufficiently fast rate, then the “borrowing of strength” from the  $m \times n$  data set  $Y_{[n]}^{[m]}$  to the inference about the data set  $Y_{[\tilde{n}]}^0$  has a striking effect: In particular, if  $f$  is an ordinary smooth kernel density, we obtain  $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/2}$ . If  $f$  is a supersmooth kernel density with smoothness  $\beta > 0$ , then  $\delta_{m,n,\tilde{n}} \asymp (1/\tilde{n})^{1/(\beta+2)}$ . (The formal definition of smoothness conditions is given in Section 2). These present sharp improvements from nonparametric rate  $(\log \tilde{n}/\tilde{n})^{1/(d+2)}$  in Eq. (6). Thus, the hierarchical models are particularly beneficial to groups of data with small sample sizes, as the convergence of the latent variable further up in the hierarchy can be translated into faster (e.g., parametric) rates of convergence of these small-sample groups. This appears to be the first result that establishes the benefits of the latent hierarchy in a concrete manner.

**Technical approach.** The major part of the proof of the main theorems lies in our attempt to establish suitable inequalities between the three quantities: (1) a Wasserstein distance

between two base measures,  $W_r(G, G')$ , (2) a suitable notion of distance between Dirichlet measures  $\mathcal{D}_{\alpha G}$  and  $\mathcal{D}_{\alpha' G'}$ , and (3) the variational distance or Kullback-Leibler divergence between the densities of  $n$ -vector  $Y_{[n]}$ , which are obtained by integrating out the (latent) Dirichlet process  $Q$  that is distributed by Dirichlet measures  $\mathcal{D}_{\alpha G}$  and  $\mathcal{D}_{\alpha' G'}$ . In fact, the establishment of these inequalities takes up the most space of this paper (Sections 3, 4 and 5). To this end, we define a notion of optimal transport distance between Dirichlet measures  $\mathcal{D}_{\alpha G}$  and  $\mathcal{D}_{\alpha' G'}$  (see (19)), which is the optimal cost of moving the mass of atoms lying in the support of measure  $\mathcal{D}_{\alpha G}$  to that of  $\mathcal{D}_{\alpha' G'}$ , where the cost of moving from an atom (that is, a measure)  $P_1 \in \mathcal{P}(\Theta)$  to another measure  $P_2 \in \mathcal{P}(\Theta)$  is again defined as a Wasserstein distance  $W_r(P_1, P_2)$  given by Eq. (3). In general, one can define distances of measures of measures and so on in a recursive way. This provides means for comparing between Bayesian hierarchies of random measures for an arbitrary number of hierarchy levels (see Section 3).

In order to derive inequalities for the aforementioned distances, our approach boils down to establishing the existence of a subset of  $\mathcal{P}(\Theta)$  which can be used to distinguish one Dirichlet measure from a class of Dirichlet measures. Because we do not have direct access to the samples  $Q_i$ 's of a Dirichlet measure, only the estimates of such samples, the test set has to be robust. By robustness, we require that the measure of a tube-set constructed along the boundary of the test set be *regular*, which means that it is possible to control the rate at which such measure vanishes, as the radius in Wasserstein metric of such tube-set tends to zero. Interestingly, the precise rates are closely linked to the geometrically sparse structure of the support of the true Dirichlet base measure. These results are developed in Section 4 and Section 5.

The proof of Theorem 2.1 requires results concerning the geometry of the support of a single Dirichlet measure. Although the support of a Dirichlet measure is very large, i.e., the entire space  $\mathcal{P}(\Theta)$  (cf. Ferguson [1973]), we show that most of the mass of a Dirichlet measure concentrates on a very small set as measured by the covering number of Wasserstein balls defined on  $\mathcal{P}(\mathbb{R}^d)$ . Our result generalizes to higher dimensions the behavior of tail probabilities chosen from a Dirichlet measure on  $\mathcal{P}(\mathbb{R})$  [Doss and Sellke, 1982].

**Organization of the paper.** Section 2 describes the model setting and provides a full statement of the main theorems. Subsection 2.3 elaborates on the components of the proofs and the tools that we develop. Section 3 defines transportation distances for hierarchies of random measures. Section 4 analyzes regular boundaries of test sets that arise in the support of various classes of Dirichlet measures of interest. Section 5 gives upper bounds for Wasserstein distances of base measures. The proof of Theorem 2.1 is given in Section 6, which draws from the machinery developed in Sections 3, 4 and 5. The proof of Theorem 2.2 is given in Section 7, which also draws on the results on the geometry of the support of a single Dirichlet measure.

**Notations.**  $W_r$  denotes the  $L_r$  Wasserstein distance.  $N(\epsilon, \mathcal{G}, W_r)$  denotes the covering number of  $\mathcal{G}$  in metric  $W_r$ .  $D(\epsilon, \mathcal{G}, W_r)$  is the packing number of the same metric [van der Vaart and Wellner, 1996].  $\text{spt } G$  denotes the support of probability measure  $G$ . Several divergence functionals of probability densities are employed:  $K(p, q)$ ,  $h(p, q)$ ,  $V(p, q)$  denote the Kullback-Leibler divergence, Hellinger and variational distance between two densities  $p$  and  $q$  defined with respect to a measure on a common space:  $K(p, q) = \int p \log(p/q)$ ,  $h^2(p, q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2$  and  $V(P, Q) = \frac{1}{2} \int |p - q|$ . In addition, we define  $K_2(p, q) = \int p [\log(p/q)]^2$ ,  $\chi(p, q) = \int p^2/q$ .  $A \lesssim B$  means  $A \leq C \times B$  for some positive constant  $C$  that is either universal or specified otherwise. Similarly for  $A \gtrsim B$ .

## 2 Main theorems and technical tools

### 2.1 Model setting and definitions

Consider the following probability model:

$$G \sim \mathcal{D}_{\gamma H}, \quad Q_1, \dots, Q_m | G \stackrel{iid}{\sim} \mathcal{D}_{\alpha G} \quad (8)$$

$$Y_{[n]}^i := (Y_{i1}, \dots, Y_{in}) | Q_i \stackrel{iid}{\sim} Q_i * f \text{ for } i = 1, \dots, m. \quad (9)$$

The relationship among quantities of interest can be illustrated by the following diagram:

$$\begin{array}{ccccc} \mathcal{D}_{\gamma H} & \longrightarrow & G & & \\ & & \downarrow & & \\ & & \mathcal{D}_{\alpha G} & & \\ & \swarrow & \downarrow & \searrow & \\ & Q_1 & \dots & Q_m & \\ & \downarrow & \downarrow & \downarrow & \\ Y_{[n]}^1 \sim Q_1 * f & \dots & Y_{[n]}^m \sim Q_m * f. & & \end{array}$$

Dropping the index  $i$ ,  $Y_{[n]} := (Y_1, \dots, Y_n)$  denotes the generic iid random  $n$ -vector according to the generic mixture density  $Q * f$ , where  $Q$  is sampled from Dirichlet measure  $\mathcal{D}_{\alpha G}$ . The marginal density of  $Y_{[n]}$  takes the form:

$$p_{Y_{[n]}|G}(Y_{[n]}) = \int \prod_{j=1}^n Q * f(Y_j) \mathcal{D}_{\alpha G}(dQ). \quad (10)$$

Given an  $m \times n$  data set  $Y_{[n]}^{[m]} := (Y_{[n]}^1, \dots, Y_{[n]}^m)$ , the posterior distribution of  $G$  given  $Y_{[n]}^{[m]}$  takes the form, for any measurable  $\mathcal{B} \subset \mathcal{P}(\Theta)$ :

$$\Pi_G(G \in \mathcal{B} | Y_{[n]}^{[m]}) = \frac{\int_{\mathcal{B}} \prod_{i=1}^m p_{Y_{[n]}|G}(Y_{[n]}^i) \mathcal{D}_{\gamma H}(dG)}{\int \prod_{i=1}^m p_{Y_{[n]}|G}(Y_{[n]}^i) \mathcal{D}_{\gamma H}(dG)} \quad (11)$$



There are two main theorems. The first is concerned with the concentration behavior of the posterior distribution of  $G$  given the data  $Y_{[n]}^{[m]}$ , as  $m, n \rightarrow \infty$ , assuming that the data is generated according to  $G = G_0$  for some fixed  $G_0 \in \mathcal{P}(\Theta)$ . The second is concerned with the concentration behavior of an individual mixing measure  $Q_i$  given the data.

**Geometric sparseness conditions for  $G_0$ .** Our theory is developed for a class of atomic base measure  $G_0$ . A simple example is the case  $G_0$  has a finite number of support points. We also consider the case  $G_0$  has infinite support, which admits a geometrically sparse structure that we now define.

**Definition 2.1.** Given  $c_1 \in (0, 1)$ ,  $c_2 > 0$  and a non-increasing function  $K : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ . A subset  $S$  of metric space  $\Theta$  is  $(c_1, c_2, K)$ -sparse if for any sufficiently small  $\delta > 0$  there is  $\epsilon \in (c_1\delta, \delta)$  according to which  $S$  can be covered by at most  $K(\epsilon)$  closed balls of radius  $\epsilon$ , and every pair of such balls is separated by a distance at least  $c_2\epsilon$ .

Probability measure  $G_0$  is said to be sparse, if its support is a  $(c_1, c_2, K)$ -sparse for a valid combination of  $c_1, c_2$  and  $K$ . A *gauge function* for a sparse measure  $G_0$ , denoted by  $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ , is defined as the maximal function such that for each sufficiently small  $\epsilon$ , there is a valid  $\epsilon$ -covering specified by the definition and that the  $G_0$  measure on each of the covering  $\epsilon$ -balls is bounded from below by  $g(\epsilon)$ .  $g$  is clearly a non-decreasing function.

We say  $G_0$  is *supersparse* with non-negative parameters  $(\gamma_0, \gamma_1)$ , if function  $K$  satisfies  $K(\epsilon) \lesssim [\log(1/\epsilon)]^{\gamma_0}$ , and function  $g$  satisfies  $g(\epsilon) \gtrsim [\log(1/\epsilon)]^{-\gamma_1}$ .  $G_0$  is *ordinary sparse* with parameters  $(\gamma_0, \gamma_1)$  if  $K(\epsilon) \lesssim (1/\epsilon)^{\gamma_0}$ , and  $g(\epsilon) \gtrsim \epsilon^{\gamma_1}$ .

**Examples.** If  $\Theta = [0, 1]$  and  $S = \{1/2^k | k \in \mathbb{N}, k \geq 1\} \cup \{0\}$ , then  $S$  is  $(c_1, c_2, K)$ -sparse with  $c_1 = 1/2, c_2 = 2$  and  $K(\epsilon) = \log(1/2\epsilon)/\log 2$ . If  $S$  is the support of  $G_0$ , and  $G_0(\{1/2^k\}) \propto k^{-\gamma_1}$  for any  $k \in \mathbb{N}$  and some  $\gamma_1 > 1$ , then  $G_0$  is clearly a supersparse measure with parameters  $\gamma_0 = 1$  and  $\gamma_1$ . Ordinary sparse measures as we defined typically arise in fractal geometry [Falconer, 1985], where parameter  $\gamma_0$  is analogous to the Hausdorff dimension of a set, while  $\gamma_1$  is analogous to the packing dimension (see, e.g. [Garcia et al., 2007]). For example, if  $\Theta = [0, 1]$  and  $S$  is the classical Cantor set, then  $S$  is  $(c, K)$ -sparse with  $c_1 = 1/3, c_2 = 2$  and  $K(\epsilon) = \exp[\log(1/2\epsilon) \log 2 / \log 3]$ . Set  $S$  has Hausdorff dimension equal  $\gamma_0 = \log 2 / \log 3$ . Let  $G_0$  be the  $\gamma_0$ -dimension Hausdorff measure on set  $S$ , then  $G_0$  is ordinary sparse with  $\gamma_0 = \gamma_1 = \log 2 / \log 3$ .

**Conditions on kernel  $f$ .** The main theorems are established independently of the specific choices of kernel density  $f$  except some minor assumptions (A1,A2) in the sequel. However, to obtain concrete rates in  $m$  and  $n$ , we will make additional assumptions on the smoothness of  $f$  when needed. Such assumptions are chosen mainly so we can make use of the rate of demixing in a deconvolution problem, i.e., the convergence rate of a point estimate of a mixing measure  $Q$  given an iid sample from the mixture density  $Q * f$ .

For that purpose,  $f$  is a density function on  $\mathbb{R}^d$  that is symmetric around 0, i.e.,  $f(x|\theta) := f(x - \theta)$  such that  $\int_B f(x)dx = \int_{-B} f(x)dx$  for any Borel set  $B \subset \mathbb{R}^d$ . In addition, the

Fourier transform of  $f$  satisfies  $\tilde{f}(\omega) \neq 0$  for all  $\omega \in \mathbb{R}^d$ . We say  $f$  is *ordinary smooth* with parameter  $\beta > 0$  if  $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}(\omega)^{-2} d\omega \lesssim (1/\delta)^{2d\beta}$  as  $\delta \rightarrow 0$ . Say  $f$  is *supersmooth* with parameter  $\beta > 0$  if  $\int_{[-1/\delta, 1/\delta]^d} \tilde{f}(\omega)^{-2} d\omega \lesssim \exp(2d\delta^{-\beta})$  as  $\delta \rightarrow 0$ . These definitions are somewhat simpler and more general than what is employed in Nguyen [2013a], who adapted from Fan [1991] to the multivariate cases.

## 2.2 Main theorems

The following list of assumptions are required throughout the paper:

- (A1) For some  $r \geq 1, C_1 > 0$ ,  $h(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1 \|\theta - \theta'\|^r$  and  $K(f(\cdot|\theta), f(\cdot|\theta')) \leq C_1 \|\theta - \theta'\|^r \forall \theta, \theta' \in \Theta$ .
- (A2) There holds  $M = \sup_{\theta, \theta' \in \Theta} \chi(f(\cdot|\theta), f(\cdot|\theta')) < \infty$ .
- (A3)  $H \in \mathcal{P}(\Theta)$  is non-atomic, and for some constant  $\eta_0 > 0$ ,  $H(B) \geq \eta_0 \epsilon^d$  for any closed ball  $B$  of radius  $\epsilon$ .

Let  $(\epsilon_n, \delta_n)_{n \geq 1}$  be two non-negative vanishing sequences, such that  $\exp -n\epsilon_n^2 = o(\delta_n)$  and that the following holds: for any  $Q \in \mathcal{P}(\Theta)$ , there exists a point estimate  $\hat{Q}_n$  given an  $n$ -iid sample from the mixture distribution  $Q * f$ , such that the following inequality holds:

$$\mathbb{P}(W_2(\hat{Q}_n, Q) \geq \delta_n) \leq 5 \exp(-cn\epsilon_n^2), \quad (12)$$

where constant  $c$  is universal, the probability measure  $\mathbb{P}$  is given by the mixture density  $Q * f$ . We refer to  $\delta_n$  as the demixing rate. The exact nature of  $(\epsilon_n, \delta_n)$  is not of concern at this point. In addition, define

$$\alpha^* := \alpha \inf_{\theta \in \text{spt } G_0} G_0(\{\theta\}).$$

Note that  $\alpha^* > 0$  if  $G$  has finite support, and  $\alpha^* = 0$  otherwise.

**Theorem 2.1.** *Let  $\Theta$  be a bounded subset of  $\mathbb{R}^d$  and  $G_0 \in \mathcal{P}(\Theta)$ . Given Assumptions (A1–A3), parameters  $\alpha \in (0, 1], \gamma > 0$  and  $H \in \mathcal{P}(\Theta)$  are known. Then, as  $n \rightarrow \infty$  and  $m = m(n) \rightarrow \infty$  there is a sequence  $\epsilon_{mn}$  dependent on  $m$  and  $n$  such that under the model given Eqs. (8) and (9), there holds:*

$$\Pi_G \left( W_1(G, G_0) \geq C\epsilon_{mn} \middle| Y_{[n]}^{[m]} \right) \rightarrow 0$$

in  $P_{Y_{[n]}|G_0}^m$ -probability for a large constant  $C$ . Moreover,

- (i) If  $G_0$  has finite (but unknown) number of support points, then

$$\epsilon_{mn} \asymp [n^d \log(mn)/m]^{1/(2d+2)} + \delta_n^{\alpha^*/(\alpha^*+1)}.$$

(ii) If  $G_0$  has infinite and supersparse support with parameters  $(\gamma_0, \gamma_1)$ , then

$$\epsilon_{mn} \asymp [n^d \log(mn)/m]^{1/(2d+2)} + \exp - [\log(1/\delta_n)]^{1/(1 \vee \gamma_0 + \gamma_1)}.$$

(iii) If  $G_0$  has infinite and ordinary sparse support with parameters  $(\gamma_0, \gamma_1)$ , then

$$\epsilon_{mn} \asymp [n^d \log(mn)/m]^{1/(2d+2)} + [\log(1/\delta_n)]^{-1/(\gamma_0 + \gamma_1)}.$$

**Remark 2.1.** Section 5 establishes the existence of a point estimate which admits the finite-sample probability bound (12). In particular,  $\epsilon_n$  is given as follows:  $\epsilon_n \asymp (\log n/n)^{r/2d}$ , if  $d > 2r$ ;  $\epsilon_n \asymp (\log n/n)^{r/(d+2r)}$  if  $d < 2r$ , and  $\epsilon_n \asymp (\log n)^{3/4}/n^{1/4}$  if  $d = 2r$ . Constant  $r$  is from Assumption A1. The rate of demixing  $\delta_n$  is determined according to an additional condition on the smoothness of the kernel density  $f$ :

- (a) If  $f$  is ordinary smooth with parameter  $\beta > 0$ , then  $\delta_n = \epsilon_n^{\frac{1}{2+\beta d'}}$  for any  $d' > d$ .
- (b) If  $f$  is supersmooth with parameter  $\beta > 0$ , then  $\delta_n = [-\log \epsilon_n]^{-1/\beta}$ .

**Remark 2.2.** If the number of support points of  $G_0$  is  $k < \infty$  and  $k$  is known;  $H$  is taken to be a probability measure with  $k$  support points. Then the proof can be easily modified to establish a parametric rate of posterior concentration:

$$\epsilon_{mn} \asymp [\log(mn)/m]^{1/2} + \delta_n^{\alpha^*},$$

Under certain strong identifiability condition for kernel density  $f$ , such as those considered by Nguyen [2013a] (Theorem 1), one has  $\epsilon_n = (\log n)n^{-1/2}$  and  $\delta_n = \epsilon_n^{1/2} = (\log n)^{1/2}n^{-1/4}$ .

**Remark 2.3.** Note that the posterior concentration rate is generally made of two terms,  $\epsilon_{mn} = A_1 + A_2$ , where  $A_1 \asymp [n^d \log(mn)/m]^{1/(2d+2)}$  vanishes as  $m \gg n^d \log n$ , and  $A_2$  vanishes as  $n \rightarrow \infty$ . In particular,  $A_2$  is defined in terms of  $\delta_n$ , the rate of demixing. It is natural to expect that  $A_2 \gg \delta_n$ , to account for the fact that the mixing measures  $Q_i$ 's are not observed directly. It is interesting how quantity  $A_2$  depends on the geometric sparsity of the support of the true base measure  $G_0$ : as  $G_0$  becomes less sparse,  $A_2$  gets slower:

$$\delta_n \ll \delta_n^{\alpha^*} \ll \delta_n^{\alpha^*/(\alpha^*+1)} \ll \exp - [\log(1/\delta_n)]^{1/(1 \vee \gamma_0 + \gamma_1)} \ll [\log(1/\delta_n)]^{-1/(\gamma_0 + \gamma_1)}.$$

**Remark 2.4.** A notable feature is the appearance of  $n$  in the nominator of quantity  $A_1 \asymp [n^d \log(mn)/m]^{1/(2d+2)}$ . One way to make sense of this is to view our  $m \times n$  asymptotic setting as that of an iid  $m$ -sample, where  $n$  plays the role of the increasing dimensionality in each of the  $m$  data point  $Y_{[n]}^i$ , for  $i = 1, \dots, m$ . [We should also point out the obvious: as  $n \rightarrow \infty$ , one obtains better estimate of the mixing measures  $Q_1, \dots, Q_m$ , which may be viewed as the  $m$ -data sample generated from the Dirichlet measure  $\mathcal{D}_{\alpha G_0}$ . Indeed, this aspect of  $n$  is already captured by the term  $A_2$  in the rate, as discussed in the previous remark]. The “ $n$  as dimensionality” viewpoint is appropriate in the sense that as  $n$  increases,

one also enlarges the space of the induced density functions  $\{P_{Y_{[n]}|G} | G \in \mathcal{P}(\Theta)\}$ , obtained by integrating out the latent Dirichlet process  $Q$  (cf. Eq. (10)). The amplification effect can be substantial, as  $G$  varies in a rather large space, the unrestricted space of measures  $\mathcal{P}(\Theta)$ . More technically, the quantity  $n^d \log n$  arises from the upper bound of Kullback-Leibler divergence  $K(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'})$ , for any pair  $(G, G') \in \mathcal{P}(\Theta)$ . The upper bound is shown to grow linearly with  $n$  (see Lemma 3.2). It is natural to expect that  $K(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'})$  increases with  $n$ , even if a tighter bound may be obtained if more is known about  $G, G'$  and  $f$ . The quantity  $n^d \log n$  also arises from the calculation of the entropy number of certain subsets of  $\mathcal{P}(\Theta)$ .

**Remark 2.5.** A more statistical explanation on the role of  $n$  in both quantities  $A_1$  and  $A_2$  may be expressed in terms of a “bias vs variance” tradeoff. Since one does not have direct observations of  $Q_1, \dots, Q_m$ , which have to be inferred through the  $n$ -samples  $Y_{[n]}^1, \dots, Y_{[n]}^m$ , respectively, increasing  $n$  results in a reduction of variance of the estimation of the  $Q_i$ ’s. On the other hand, increasing  $n$  also results in an increased bias placed on the estimates of the specific  $m$ -sample  $Q_1, \dots, Q_m$ , based upon which one makes inference about the base measure  $G$  living in a large space. Thus, it is conceivable that for a given choice of  $n$ , instead of increasing  $n$  further, one may be better off increase  $m$  to improve the inference. Indeed, our bound suggests there is an optimal regime at which  $m$  increases relatively to  $n$ , as both tend to infinity.

**Remark 2.6.** Returning to Remark 2.2, in parametric models, it is not uncommon to encounter the appearance of convergence rates (e.g., for covariance estimation, or sparse regression) that grow with  $\log(n)/m$ , when  $m$  is the sample size, and  $n$  defines the dimensionality of the sample. See, e.g., Buhlmann and van de Geer [2011] for numerous examples in high-dimensional parametric inference. It is interesting to note that as one goes from parametric to nonparametric models, due to the infinite dimensionality of  $Q_i$ ’s, the complexity term associated with  $n$  grows from  $\log n$  to a polynomial in  $n$ .

The next result is about the posterior concentration behavior of the latent mixing measures  $Q_i$ ’s, as the base measure  $G$  is integrated out, and the amount of data increases. For the ease of presentation, we isolate a particular mixing measure to be denoted by  $Q_0$ , and we shall assume that  $Q_0$  is attached to the hierarchical Dirichlet process in the same way as the  $Q_1, \dots, Q_m$ , i.e.:

$$G \sim \mathcal{D}_{\gamma H}, \quad Q_0, Q_1, \dots, Q_m | G \stackrel{iid}{\sim} \mathcal{D}_{\alpha G}. \quad (13)$$

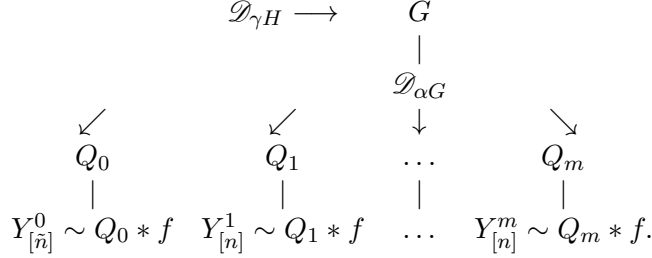
Suppose that an iid  $\tilde{n}$ -sample  $Y_{[\tilde{n}]}^0$  drawn from a mixture model  $Q_0 * f$  is available, where  $Q_0 = Q_0^* \in \mathcal{P}(\Theta)$  is unknown:

$$Y_{[\tilde{n}]}^0 | Q_0 \stackrel{iid}{\sim} Q_0 * f. \quad (14)$$

In addition, as before,  $m \times n$  data set is available:

$$Y_{[n]}^i := (Y_{i1}, \dots, Y_{in}) | Q_i \stackrel{iid}{\sim} Q_i * f \text{ for } i = 1, \dots, m. \quad (15)$$

The relationship among quantities of interest is illustrated by the following diagram:



The following theorem shows that the posterior distribution  $\Pi(Q_0|Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]})$ , defined with respect to specifications (13), (14) and (15), concentrates most its mass toward  $Q_0^*$ , as  $n, m$  and  $\tilde{n} \rightarrow \infty$  appropriately. Motivated by the conclusion of Theorem 2.1 we shall assume that the posterior distribution of  $G$  concentrates at a certain rate  $\epsilon_{mn}$  toward the true base measure  $G_0$ , which is now assumed to have a finite (but unknown) number of support points. This concentration behavior can in turn be translated to a sharp concentration behavior for the mixing measure  $Q_0$ . A complete statement of the theorem is as follows:

**Theorem 2.2.** *Let  $\Theta$  be a bounded subset of  $\mathbb{R}^d$ ,  $G_0, Q_0^* \in \mathcal{P}(\Theta)$ . Suppose that Assumptions (A1) and (A2) hold for some  $r \geq 1$ . Given parameters  $\alpha \in (0, 1]$ ,  $\gamma > 0$ , and  $H \in \mathcal{P}(\Theta)$  are known. Assume further that*

- (a)  $G_0$  has  $k < \infty$  support points in  $\Theta$ ;  $Q_0^* \in \mathcal{P}(\Theta)$  such that  $\text{spt } Q_0^* \subseteq \text{spt } G_0$ .
- (b) For each  $\tilde{n}$ , there is a net  $\epsilon_{mn} = \epsilon_{mn}(\tilde{n}) \downarrow 0$  indexed by  $m, n$  such that under the model specifications (13), (14) and (15), there holds:

$$\Pi_G \left( W_1(G, G_0) \geq C\epsilon_{mn} \middle| Y_{[n]}^{[m]}, Y_{[\tilde{n}]}^0 \right) \longrightarrow 0$$

in  $P_{Y_{[n]}^0|G_0}^m \times P_{Y_{[\tilde{n}]}^0|Q_0^*}$ -probability, as  $n \rightarrow \infty$  and  $m = m(n) \rightarrow \infty$  at a suitable rate with respect to  $n$ . Here,  $C$  is a constant independent of  $\tilde{n}, m, n$ .

Then, as  $\tilde{n} \rightarrow \infty$  and then  $n$  and  $m = m(n) \rightarrow \infty$ , we have

$$\Pi_Q \left( h(Q_0 * f, Q_0^* * f) \geq \delta_{m,n,\tilde{n}} \middle| Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]} \right) \longrightarrow 0$$

in  $P_{Y_{[\tilde{n}]}^0|Q_0^*} \times P_{Y_{[n]}^0|G_0}^m$ -probability, where the rates  $\delta_{m,n,\tilde{n}}$  are given as follows:

- (i)  $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/(d+2)} + \epsilon_{mn}^{r/2} \log(1/\epsilon_{mn})$ .
- (ii)  $\delta_{m,n,\tilde{n}} \asymp (\log \tilde{n}/\tilde{n})^{1/2}$  if  $f$  is ordinary smooth with smoothness  $\beta > 0$ , and  $n$  and  $m$  grow sufficiently fast so that  $\epsilon_{mn} \lesssim \tilde{n}^{-(\alpha+k+M_0)} (\log \tilde{n})^{-(\alpha+k-2)}$  for some constant  $M_0 > 0$  depending only on  $d, k, \beta$  and  $\text{diam}(\Theta)$ .

(iii)  $\delta_{m,n,\tilde{n}} \asymp (1/\tilde{n})^{1/(\beta+2)}$ , if  $f$  is supersmooth with smoothness  $\beta > 0$ ,  $n$  and  $m$  grow sufficiently fast so that  $\epsilon_{mn} \lesssim \tilde{n}^{-2(\alpha+k)/(\beta+2)} (\log \tilde{n})^{-2(\alpha+k-1)} \exp(-4\tilde{n}^{\beta/(\beta+2)})$ .

**Remark 2.7.** Condition (a) that  $\text{spt } Q_0^* \subset \text{spt } G_0$  motivates the incorporation of mixture distribution  $Q_0 * f$  to the Bayesian hierarchy as specified by Eq. (13). According to the model,  $Q_0$  shares the same supporting atoms with  $Q_1, \dots, Q_m$ , as they all inherit from random base measure  $G$ . Note also that the condition of posterior of  $G$  as stated in (b) is closely related to but nonetheless different from the conclusion reached by Theorem 2.1, due to the additional conditioning on  $Y_{[\tilde{n}]}^0$ . This condition may be proved directly under additional conditions on  $Q_0^*$  and  $G_0$ , by a technically cumbersome (but conceptually simple) modification of the proof of Theorem 2.1. We avoid this unnecessary complication as it is not central to the main message of the present theorem.

**Remark 2.8.** In a stand-alone setting  $Q_0$  is endowed with a Dirichlet prior:  $Q_0 \sim \mathcal{D}_{\alpha_0 H_0}$  for some known  $\alpha_0 > 0$  and non-atomic base measure  $H_0 \in \mathcal{P}(\Theta)$ . Combining with the model specification expressed by (14), we obtain the posterior distribution for  $Q_0$ , which admits the following concentration behavior under some mild condition (cf. Nguyen [2013a]):

$$\Pi_Q \left( h(Q_0 * f, Q_0^* * f) \geq (\log \tilde{n}/\tilde{n})^{\frac{1}{d+2}} \middle| Y_{[\tilde{n}]}^0 \right) \rightarrow 0 \quad (16)$$

in  $P_{Y_{[\tilde{n}]}^0 | Q_0^*}$ -probability. This should be compared to the general concentration rate given by claim (i) of Theorem 2.2:  $(\log \tilde{n}/\tilde{n})^{1/(d+2)} + \epsilon_{mn}^{r/2} \log(1/\epsilon_{mn})$ . The extra quantity  $\epsilon_{mn}^{r/2} \log(1/\epsilon_{mn})$  can be viewed as the overhead cost for maintaining the latent hierarchy involving the random Dirichlet prior  $\mathcal{D}_{\alpha G}$ .

**Remark 2.9.** Claims (ii) and (iii) demonstrate the benefits of hierarchical modeling for groups of data with relatively small sample size: when  $n \gg \tilde{n}$  (and  $m = m(n) \rightarrow \infty$  suitably) so that  $\epsilon_{mn}$  is sufficiently small, we obtain parametric rates for the mixture density  $Q_0 * f$ :  $(\log \tilde{n}/\tilde{n})^{1/2}$  for ordinary smooth kernels, and  $(1/\tilde{n})^{1/(\beta+2)}$  for supersmooth kernels. This is a sharp improvement over the standard rate  $(\log \tilde{n}/\tilde{n})^{1/(d+2)}$  one would get for fitting a stand-alone mixture model  $Q_0 * f$  using a Dirichlet process prior. This improvement is possible due to the confluence of two factors: By attaching  $Q_0$  to the Bayesian hierarchy one is able to exploit the assumption that random measure  $Q_0$  shares the same supporting atoms as the random base measure  $G$ . This is translated to a favorable level of thickness of the conditional prior for  $Q_0$  (given the  $m \times n$  data  $Y_{[n]}^{[m]}$ ), as measured by small Kullback-Leibler neighborhoods. The second factor is due to our new construction of a sieves (subsets of)  $\mathcal{P}(\Theta)$  over which the Dirichlet measure concentrates most its mass on, but which have suitably small entropy numbers. These details will be elaborated in Section 7.

In summary, Theorem 2.1 establishes posterior concentration of the Dirichlet base measure in a hierarchical setting, while Theorem 2.2 demonstrates dramatic gains in efficiency in the inference of groups of data with relatively small sample size. For groups with relatively large sample size, the concentrate rate is weakened. This quantifies the effects of

“borrowing of strength”, from large groups of data to smaller groups. This is arguably a good virtue of hierarchical models: it is the populations with small sample sizes that need improved inference the most.

### 2.3 Method of proof

The major part of the proof of Theorem 2.1 lies in our attempt to establish the relationship (in inequalities) between the three important quantities: (1) a Wasserstein distance between two base measures,  $W_r(G, G')$ , (2) a suitable notion of distance between Dirichlet measures  $\mathcal{D}_{\alpha G}$  and  $\mathcal{D}_{\alpha G'}$ , and (3) the variational distance/ Kullback-Leibler divergence between the marginal densities of  $n$ -vector  $Y_{[n]}$ , which are obtained by integrating out the mixing measure  $Q$ , which is a Dirichlet process distributed by  $\mathcal{D}_{\alpha G}$  and  $\mathcal{D}_{\alpha G'}$ , respectively. The link from  $G$  (resp.  $G'$ ) to the induced  $P_{Y_{[n]}|G}$  (resp.  $P_{Y_{[n]}|G'}$ ) is illustrated by the following diagram:

$$\left\{ \begin{array}{ccccccc} G & \rightarrow & \mathcal{D}_{\alpha G} & \rightarrow & Q & \rightarrow & Q * f \rightarrow Y_{[n]}. \\ | & & | & & & & | \\ W_r(G, G') & & W_r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G'}) & & & & V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \\ | & & | & & & & | \\ G' & \rightarrow & \mathcal{D}_{\alpha G'} & \rightarrow & Q & \rightarrow & Q * f \rightarrow Y_{[n]}. \end{array} \right.$$

In order to establish the relationship among the aforementioned distances we need to investigate the geometry of the support of the individual Dirichlet measures, as well as the geometry of test sets that arise when a given Dirichlet measure is tested (discriminated) from a large class of Dirichlet measures. In fact, this study forms the bulk of the paper in Section 3, Section 4, and Section 5.

**Transportation distances for Bayesian hierarchies.** To begin, in Section 3 we develop a general notion of transportation distance of Bayesian hierarchies of random measures. This notion plays a fundamental role in our theory, and we believe is also of independent interest. Using transportation distances it is possible to compare between not only two probability measures defined on  $\Theta$ , but also two probability measures on the space of measures on  $\Theta$ , and so on. Transportation distances are natural for comparing between Bayesian hierarchies, because the geometry of the space of support of measures is inherited directly into the definition of the transportation distances between the measures. In particular,  $W_r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G'})$  is defined as the Wasserstein distance on the Polish space  $\mathcal{P}(\mathcal{P}(\Theta))$ , by inheriting the Wasserstein distance on the Polish space of measures  $\mathcal{P}(\Theta)$ . (The notation  $W_r$  is reused as a harmless abuse of notation). It can be shown that

$$W_r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G'}) \geq W_r(G, G').$$

The above inequalities hold generally if  $\mathcal{D}_{\alpha G}$  and  $\mathcal{D}_{\alpha' G'}$  are replaced by any pair of probability measures on  $\mathcal{P}(\Theta)$  that admit a suitable notion of mean measures  $G$ , and  $G'$ , respectively. Moreover, the Dirichlet measures allow a remarkable identity: when  $\alpha = \alpha'$ , there also holds that

$$W_r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G'}) = W_r(G, G').$$

A series of application of Jensen's inequality yields the following upper bound for the KL divergence:<sup>3</sup>

$$h^2(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \leq K(P_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \lesssim nW_r^r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G'}) = nW_r^r(G, G').$$

**Bounds on Wasserstein distances.** The most demanding part of the paper lies in establishing an upper bound of the Wasserstein distance  $W_r(G, G')$  in terms of the variational distance  $V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$ . This is achieved by Theorem 5.1 in Section 5, which states that for a fixed  $G \in \mathcal{P}(\Theta)$  and any  $G' \in \mathcal{P}(\Theta)$ ,

$$W_r^r(G, G') \lesssim V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) + A_n(G, G'), \quad (17)$$

where  $A_n(G, G')$  is a quantity that tends to 0 as  $n \rightarrow \infty$ . The rate at which  $A_n(G, G')$  tends zero depends only on the geometrically sparse structure of  $G$ , not  $G'$ . The proof of this result hinges on the existence of a suitable set  $\mathcal{B}_n \subset \mathcal{P}(\Theta)$  measurable with respect to (the sigma algebra induced by) the observed variables  $Y_{[n]}$ , which can then be used to distinguish  $G'$  from  $G$ , in the sense that

$$W_r^r(G, G') \lesssim P_{Y_{[n]}|G'}(\mathcal{B}_n) - P_{Y_{[n]}|G}(\mathcal{B}_n) + A_n(G, G'). \quad (18)$$

We develop two main lines of attack to arrive at a construction of  $\mathcal{B}_n$ .

First, we establish the existence of a point estimate for the mixing measure on the basis of the observed  $Y_{[n]}$ . Moreover, such point estimates have to admit a finite-sample probability bound of the following form: given  $Y_{[n]} \sim Q * f$ , there exist a point estimate  $\hat{Q}_n$  such that under the  $Q * f$  probability, there holds

$$\mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta_n) \leq \exp -n\epsilon_n^2,$$

where  $\delta_n$  and  $\epsilon_n$  are suitable vanishing sequences. These finite-sample bounds are presented in Section 5. The existence of  $\hat{Q}_n$  will then be utilized in the construction of a suitable set  $\mathcal{B}_n$ . In particular, one may pretend to have direct observations from the Dirichlet measures to construct the test sets, with a possible loss of accuracy captured by the demixing rate  $\delta_n$ .

**Regular boundaries in the support of Dirichet measures.** Now, to control  $A_n(G, G')$ , we need the second piece of the argument, which establishes the existence of a robust test that can be used to distinguish a Dirichlet measure  $\mathcal{D}_{\alpha G}$  from a class of Dirichet measures

---

<sup>3</sup>Within this subsection, the details on the constants underlying  $\lesssim$  and  $\gtrsim$  are omitted for the sake of brevity.



$\mathcal{C} = \{\mathcal{D}_{\alpha'G'} | G' \in \mathcal{P}(\Theta)\}$ , where the robustness here is measured by Wasserstein metric  $W_r$  on  $\mathcal{P}(\Theta)$ . The robustness is needed to account for the possible loss of accuracy  $\delta_n$  incurred by demixing, as alluded to in the previous paragraph. A formal theory of robust tests is developed in Section 4. Central to this theory is a notion of regularity for a given class of Dirichlet measures  $\mathcal{C}$  with respect to a fixed Dirichlet measure  $\mathcal{D} := \mathcal{D}_{\alpha G}$ . In particular, we say that  $\mathcal{C}$  has regular boundary with respect to  $\mathcal{D}$  if for each element  $\mathcal{D}' = \mathcal{D}_{\alpha'G'} \in \mathcal{C}$  there is a measurable subset  $\mathcal{B} \subset \mathcal{P}(\Theta)$  for which the following holds: (i)  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) \gtrsim W_r^r(G, G')$  and (ii)

$$\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) \rightarrow 0$$

as  $\delta \rightarrow 0$ . Set  $\mathcal{B}$  can be thought of as a test set which is used to approximate the variation distance between a fixed  $\mathcal{D}$  and an arbitrary  $\mathcal{D}'$  which varies in  $\mathcal{C}$ . Various forms of regularity are developed, which specifies how fast the quantity in the previous display tends to 0. Thus, the achievement of this section is to show that the regularity behavior is closely tied to the geometry of the support of base measure  $G$ . Theorem 4.1 and Theorem 4.2 provide a complete picture of regularity for the case  $G$  has finite support, and the case  $G$  has infinite and geometrically sparse support. Now, by controlling the rate at which  $\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B})$  tends to 0, we can control the rate at which  $A_n(G, G')$  tends to 0, completing the proof of (17).

**Posterior concentration proof.** With the tools and inequalities established in Section 3, 4 and 5 at our disposal, the proof of Theorem 2.1 becomes a straightforward exercise of calculations. At a high level, the proof of posterior concentration for the base measure  $G$  follows a general strategy for analyzing the convergence of mixing measures in mixture models. This strategy owes its roots to the standard framework developed by Ghosal et al. [2000], which characterizes the posterior concentration by sufficient conditions in terms of entropy numbers, the prior thickness in Kullback-Leibler divergence, and so on. These sufficient conditions are verified using the inequalities established by earlier sections. See Section 6 for details.

**Posterior concentration under perturbed base measure of a Dirichlet prior.** Finally, the proof of Theorem 2.2 follows from a posterior concentration result for the mixing measure  $Q$ , which is distributed by the prior  $\mathcal{D}_{\alpha G}$ , conditionally given the event that the base measure  $G$  is perturbed by a small Wasserstein distance  $W_1$  from  $G_0$  that has  $k < \infty$  support points, see Lemma 7.4 in Section 7. The proof of this lemma also follows the standard strategy of the posterior concentration proof mentioned earlier. The main novelty lies in the construction of a sieves of subsets of  $\mathcal{P}(\Theta)$  which yields favorable rates of posterior concentration. This construction is possible by showing that the Dirichlet measure places most its mass on subsets (of  $\mathcal{P}(\Theta)$ ) which can be covered by a relatively small number of balls in  $W_r$ . Such results about the Wasserstein geometry of the support of a Dirichlet measure may be of independent interest, and are collected in section 7.2.

## 2.4 Concluding remarks and further development

We established posterior concentration rates for the base measure of a Dirichlet measure given observations associated with sampled Dirichlet processes, using tools developed with optimal transport distances. Motivated by the hierarchical Dirichlet processes, whose base measure is discrete, we developed asymptotic results for the case the true base measure is atomic with finite or infinite and sparse support. The problem of estimating a non-atomic base measure, while the Dirichlet processes are not directly observed, remains open. In fact, it appears that no practical estimation method for this setting exists as of this writing. Finally, our results only establish upper bounds for the posterior concentration of various latent measure-valued quantities. It is also of interest to obtain lower bounds on concentration rates, and to develop a minimax optimal theory, for the variables residing in latent hierarchies.

## 3 Transportation distances of Bayesian hierarchies

Let  $\Theta$  be a complete separable metric space (i.e.,  $\Theta$  is a Polish space) and  $\mathcal{P}(\Theta)$  be the space of Borel probability measures on  $\Theta$ . The weak topology on  $\mathcal{P}(\Theta)$  (or narrow topology) is induced by convergence against  $C_b(\Theta)$ , i.e., bounded continuous test functions on  $\Theta$ . Since  $\Theta$  is Polish,  $\mathcal{P}(\Theta)$  is itself a Polish space.  $\mathcal{P}(\Theta)$  is metrized by the  $W_r$  Wasserstein distance: for  $G, G' \in \mathcal{P}(\Theta)$  and  $r \geq 1$ ,

$$W_r(G, G') = \inf_{\kappa \in \mathcal{T}(G, G')} \left[ \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.$$

By a recursion of notations,  $\mathcal{P}^{(2)}(\Theta) := \mathcal{P}(\mathcal{P}(\Theta))$  is defined as the space of Borel probability measures on  $\mathcal{P}(\Theta)$ . This is a Polish space, and will be endowed again with a Wasserstein metric that is induced by metric  $W_r$  on  $\mathcal{P}(\Theta)$ :

$$W_r(\mathcal{D}, \mathcal{D}') = \inf_{\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')} \left[ \int W_r^r(G, G') d\mathcal{K}(G, G') \right]^{1/r}. \quad (19)$$

We can safely reuse notation  $W_r$  as the context is clear from the arguments. Since the cost function  $\|\theta - \theta'\|$  is continuous, the existence of an optimal coupling  $\kappa \in \mathcal{T}(G, G')$  which achieves the infimum is guaranteed due to the tightness of  $\mathcal{T}(G, G')$  (cf. Theorem 4.1 of Villani [2008]). Moreover,  $W_r(G, G')$  is a continuous function and  $\mathcal{T}(\mathcal{D}, \mathcal{D}')$  is again tight, so the existence of an optimal coupling in  $\mathcal{T}(\mathcal{D}, \mathcal{D}')$  is also guaranteed.

Now we present a lemma on a monotonic property of Wasserstein metrics defined along the recursive construction for every pair of centered random measures on  $\Theta$ . Part (b) highlights a very special property of the Dirichlet measure. In what follows  $P$  denotes a generic measure-valued random variable. By  $\int P d\mathcal{D} = G$  we mean  $\int P(A) d\mathcal{D} = G(A)$  for any measurable subset  $A \subset \Theta$ .

**Lemma 3.1.** (a) Let  $\mathcal{D}, \mathcal{D}' \in \mathcal{P}^{(2)}(\Theta)$  such that  $\int P d\mathcal{D} = G$  and  $\int P d\mathcal{D}' = G'$ . For  $r \geq 1$ , if  $W_r(\mathcal{D}, \mathcal{D}')$  is finite then  $W_r(\mathcal{D}, \mathcal{D}') \geq W_r(G, G')$ .

(b) Let  $\mathcal{D} = \mathcal{D}_{\alpha G}$  and  $\mathcal{D}' = \mathcal{D}_{\alpha G'}$ . Then,  $W_r(\mathcal{D}, \mathcal{D}') = W_r(G, G')$  if both quantities are finite.

Recall the generative process defined by Eqs (8) and (9): The marginal density  $p_{Y_{[n]}|G}$  is obtained by integrating out random measures  $Q$ , which is distributed by  $\mathcal{D}_{\alpha G}$ , see Eq (10). By a repeated application of Jensen's inequality, it is simple to establish upper bounds on Kullback-Leibler distance  $K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$  and other related distances in terms of transportation distance between  $G$  and  $G'$ .

**Lemma 3.2.** (a) Under assumption (A1),

$$\begin{aligned} K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) &\leq C_1 n W_r^r(G, G') \\ h^2(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) &\leq C_1 n W_{2r}^{2r}(G, G'). \end{aligned}$$

(b) Under assumption (A2), we have  $\chi(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq M^n$ .

Next, define the Kullback-Leibler neighborhood of a given  $G_0 \in \mathcal{P}(\Theta)$  with respect to  $n$ -vector  $Y_{[n]}$  as follows:

$$B_K(G_0, \delta) = \{G \in \mathcal{P}(\Theta) | K(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq \delta^2, K_2(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq \delta^2\}. \quad (20)$$

The following result gives probability bound on small balls as defined by Wasserstein metric (Lemma 5 of Nguyen [2013a]):

**Lemma 3.3.** Suppose that  $\text{law}(G) = \mathcal{D}_{\gamma H}$ , where  $H$  is a non-atomic probability measure on  $\Theta$ . For a small  $\epsilon > 0$ , let  $D = D(\epsilon, \Theta, \|\cdot\|)$  the packing number of  $\Theta$  under  $\|\cdot\|$ . Then, for any  $G_0 \in \mathcal{P}(\Theta)$ ,

$$\mathbb{P}\left(G : W_r^r(G_0, G) \leq (2^r + 1)\epsilon^r\right) \geq \frac{\Gamma(\gamma)\gamma^D}{(2D)^{D-1}} \left(\frac{\epsilon}{\text{diam}(\Theta)}\right)^{r(D-1)} \sup_S \prod_{i=1}^D H(S_i).$$

Here,  $(S_1, \dots, S_D)$  denotes the  $D$  disjoint  $\epsilon/2$ -balls that form a maximal packing of  $\Theta$ .  $\Gamma(\cdot)$  denotes the gamma function. The supremum is taken over all packings  $S := (S_1, \dots, S_D)$ .

Combine the previous lemmas to obtain an estimate of the thickness of the hierarchical Dirichlet prior:

**Lemma 3.4.** Given assumptions (A1–A3),  $\Theta$  a bounded subset of  $\mathbb{R}^d$ . Then, for  $D := (\text{diam}(\Theta))^d (n^3/\delta)^{d/r}$  and constants  $c, C$  depending only on  $C_1, M, \eta_0, \gamma$ ,  $\text{diam}(\Theta)$  and  $r$ , for any  $G_0 \in \mathcal{P}(\Theta)$ ,  $\delta > 0$  and  $n > C \log(1/\delta)$ , the following inequality holds under the probability measure  $\mathcal{D}_{\gamma H}$ :

$$\log \mathbb{P}(G \in B_K(G_0, \delta)) \geq c \log \left[ \gamma^D (\delta^2/n^3)^{(1+d/r)(D-1)+Dd/r} \right].$$

The proofs of all lemmas presented in this section are deferred to Section 8.

## 4 Regular boundaries in the support of Dirichlet measures

In this section we study the property of the boundary of certain sets (of measures) which can be used to test one Dirichlet measure against another. Typically such a test set can be defined via the variational distance between the two measures. However, for the purpose of subsequent development we need a more robust test in which the robustness can be expressed in terms of the measure of the test set's perturbation along its boundary. Recall the variational distance between  $\mathcal{D}, \mathcal{D}' \in \mathcal{P}^{(2)}(\Theta)$  is given by:

$$V(\mathcal{D}, \mathcal{D}') = \sup_{\mathcal{B} \subset \mathcal{P}(\Theta)} |\mathcal{D}(\mathcal{B}) - \mathcal{D}'(\mathcal{B})|.$$

Here the supremum is taken over all Borel measurable sets  $\mathcal{B} \subset \mathcal{P}(\Theta)$ . In what follows, fix  $r \geq 1$ . For a subset  $\mathcal{B} \subset \mathcal{P}(\Theta)$  the boundary set  $\text{bd } \mathcal{B}$  is defined as the set of all elements  $P \in \mathcal{P}(\Theta)$  such that every  $W_r$  neighborhood for  $P$  has non-empty intersection with  $\mathcal{B}$  as well as the complement set  $\mathcal{B}^c = \mathcal{P}(\Theta) \setminus \mathcal{B}$ . Also define  $\mathcal{B}_\epsilon$  to be the set of all  $P \in \mathcal{P}(\Theta)$  for which there is a  $Q \in \mathcal{B}$  and  $W_r(Q, P) \leq \epsilon$ .

The primary objects in consideration are a pair of  $(\mathcal{D}, \mathcal{C}) \in (\mathcal{P}(\Theta), \mathcal{P}^{(2)}(\Theta))$ , where  $\mathcal{D} = \mathcal{D}_{\alpha G}$  for some fixed  $G \in \mathcal{P}(\Theta)$  and  $\alpha > 0$ .  $\mathcal{C}$  is a class of Dirichlet measures  $\mathcal{C} := \{\mathcal{D}_{\alpha' G'} | G' \in \mathcal{G}, \alpha' > 0\}$  for some fixed  $\mathcal{G} \subset \mathcal{P}(\Theta)$ .

**Definition 4.1.** A class  $\mathcal{C} \subset \mathcal{P}^{(2)}(\Theta)$  of Dirichlet measures is said to have  $\alpha^*$ -regular boundary with respect to  $\mathcal{D} = \mathcal{D}_{\alpha G}$  for some constant  $\alpha^* > 0$ , if there are positive constants  $C_0, c_0$  and  $c_1$  dependent only on  $\mathcal{D}$  such that for each  $\mathcal{D}' = \mathcal{D}_{\alpha' G'} \in \mathcal{C}$  there exists a measurable subset  $\mathcal{B} \subset \mathcal{P}(\Theta)$  for which the following hold:

- (i)  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) \geq c_0 W_r^r(G, G')$ ,
- (ii)  $\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) \leq C_0 \left( \delta / W_r(G, G') \right)^{\alpha^*}$  for any  $\delta \leq c_1 W_r(G, G')$ .

$\mathcal{C}$  is said to have strong  $\alpha^*$ -regularity with respect to  $\mathcal{D}$  if condition (ii) is replaced by

- (iii)  $\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) \leq C_0 \delta^{\alpha^*}$  for any  $\delta \leq c_1$ .

$\mathcal{C}$  is said to have weak regularity with respect to  $\mathcal{D}$  if condition (ii) is replaced by

- (iv)  $\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) = o(1)$  as  $\delta \rightarrow 0$ .

**Remark.** The nontrivial requirement here is that constants  $C_0, c_0$  and  $c_1$  are independent of  $\mathcal{D}' \in \mathcal{C}$ . Consider the following example:  $\mathcal{G} := \{G' \in \mathcal{P}(\Theta) | \text{spt } G' \cap \text{spt } G = \emptyset\}$ . Take  $\mathcal{D}' := \mathcal{D}_{\alpha' G'}$  for some  $G' \in \mathcal{G}$ . By a standard fact of Dirichlet measures (e.g., see Thm. 3.2.4 of Ghosh and Ramamoorthi [2002]),  $\text{spt } \mathcal{D} = \{P : \text{spt } P \subset \text{spt } G\}$  and  $\text{spt } \mathcal{D}' = \{P : \text{spt } P \subset \text{spt } G'\}$ . Thus, we also have  $\text{spt } \mathcal{D} \cap \text{spt } \mathcal{D}' = \emptyset$ . It follows that  $V(\mathcal{D}, \mathcal{D}') = 1$ . If we choose  $\delta_1 = \inf_{\theta \in \text{spt } G; \theta' \in \text{spt } G'} \|\theta - \theta'\| > 0$ , and let  $\mathcal{B} = (\text{spt } \mathcal{D}')_{\delta_1/2}$ , then  $\mathcal{D}'(\mathcal{B}) = 1$  and  $\mathcal{D}(\mathcal{B}) = 0$ . Moreover, for any  $\delta \leq \delta_1/4$ ,  $\mathcal{D}(\mathcal{B}_\delta) = 0$ ,

so  $\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) = 0$ . At the first glance, this construction appears to suggest that  $\mathcal{C} := \{\mathcal{D}_{\alpha'G'} | G' \in \mathcal{G}\}$  has (strong)  $\alpha^*$ -regular boundary with  $\mathcal{D}$  for any  $\alpha^* > 0$ . This is not the case, because it is not possible to guarantee that  $\delta_1 > c_1 W_r(G, G')$  for some  $c_1$  independent of  $G'$ . That is,  $\delta_1$  can be arbitrarily close to 0 even as  $W_r(G, G')$  remains bounded away from 0.

#### 4.1 The case of finite support

We study the regularity of boundaries for the pair  $(\mathcal{D}, \mathcal{C})$ , where the base measure  $G$  of  $\mathcal{D} = \mathcal{D}_{\alpha G}$  has a finite number of support points, while class  $\mathcal{C}$  consists of Dirichlet measures  $\mathcal{D}' = \mathcal{D}_{\alpha G'}$  where  $G'$  may have infinite support in  $\Theta$ . In the following subsection we extend the theory to handle the case that  $G$  has infinite and geometrically sparse support.

**Theorem 4.1.** *Suppose that  $\Theta$  is bounded. Let  $\mathcal{D} = \mathcal{D}_{\alpha G}$ , where  $G = \sum_{i=1}^k \beta_i \delta_{\theta_i}$  for some  $k < \infty$  and  $\alpha \in (0, 1]$ . Let  $\alpha_1 > \alpha_0 > 0$  be given. Define*

$$\mathcal{C} = \{\mathcal{D}_{\alpha'G'} | G' \in \mathcal{P}(\Theta); \alpha' \in [\alpha_0, \alpha_1]\}.$$

*Then,  $\mathcal{C}$  has  $\alpha^*$ -regular boundary with respect to  $\mathcal{D}$ , where  $\alpha^* = \min_i \alpha \beta_i$ .*

*Proof.* Take any  $G' \in \mathcal{P}(\Theta)$ . Let  $\epsilon := W_r(G, G')$ . Choose constants  $c_1, c_2$  such that  $c_1^r + c_2 \text{diam}(\Theta)^r \leq 1/2^r$  and  $c_1 \text{diam}(\Theta) < m := \min_{1 \leq i \neq j \leq k} \|\theta_i - \theta_j\|/4$ . Let  $S = \cup_{i=1}^k B_i$ , where  $B_i$ 's for  $i = 1, \dots, k$  are closed Euclidean balls of radius  $c_1 \epsilon$  and centering at  $\theta_1, \dots, \theta_k$ , respectively. Any  $G' \in \mathcal{P}(\Theta)$  admits either (A)  $G'(S^c) \geq c_2 \epsilon^r$ , or (B)  $G'(S^c) < c_2 \epsilon^r$ .

**Case (A)**  $G'(S^c) \geq c_2 \epsilon^r$ . Let  $\mathcal{B} = \{Q \in \mathcal{P}(\Theta) | Q(S^c) > 1/2\}$ . Clearly,  $\mathcal{D}(\mathcal{B}) = 0$ . Moreover, for any  $Q \in \mathcal{B}$  and  $Q' \in \text{spt } \mathcal{D}$ ,  $W_r^r(Q, Q') \geq (1/2)(c_1 \epsilon)^r$ . So for any  $\delta < (1/2)^{1/r} c_1 \epsilon$ ,  $\mathcal{D}(\mathcal{B}_\delta) = 0$ . Condition (ii) of Definition 4.1 is satisfied.

It remains to verify condition (i). If  $G'(S) = 0$  then  $G'(S^c) = 1$  and  $\mathcal{D}'(\mathcal{B}) = 1$ . So,  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) = 1$ . On the other hand, if  $G'(S) > 0$  and suppose that  $\text{law}(Q) = \mathcal{D}'$ , then  $\text{law}(Q(S)) = \text{Beta}(\alpha' G'(S), \alpha' G'(S^c))$ . So,

$$\begin{aligned} \mathcal{D}'(\mathcal{B}) &= \int_0^{1/2} \frac{\Gamma(\alpha')}{\Gamma(\alpha' G'(S)) \Gamma(\alpha' G'(S^c))} x^{\alpha' G'(S)-1} (1-x)^{\alpha' G'(S^c)-1} dx \\ &\geq \frac{(1/2)^{\alpha'} \Gamma(\alpha')}{\Gamma(\alpha' G'(S)) \Gamma(\alpha' G'(S^c))} \int_0^{1/2} x^{\alpha' G'(S)-1} dx \\ &= \frac{(1/2)^{\alpha'} \Gamma(\alpha')}{\Gamma(\alpha' G'(S)) \Gamma(\alpha' G'(S^c))} \times \frac{(1/2)^{\alpha' G'(S)}}{\alpha' G'(S)} \\ &= \frac{(1/2)^{\alpha' + \alpha' G'(S)} \Gamma(\alpha') \alpha' G'(S^c)}{\Gamma(\alpha' G'(S) + 1) \Gamma(\alpha' G'(S^c) + 1)} \\ &\geq \frac{(1/2)^{2\alpha'} \Gamma(\alpha') \alpha' G'(S^c)}{\max_{1 \leq x \leq \alpha' + 1} \Gamma(x)^2} \geq \frac{(1/2)^{2\alpha'} \Gamma(\alpha') \alpha' c_2 \epsilon^r}{\max_{1 \leq x \leq \alpha' + 1} \Gamma(x)^2}. \end{aligned}$$

In the above display, the first inequality is due to  $(1-x)^\gamma \geq 1$  if  $\gamma \leq 0$ , and  $(1-x)^\gamma \geq (1/2)^\gamma$  if  $\gamma > 0$  for  $x \in [0, 1/2]$ . The third equality is due to  $x\Gamma(x) = \Gamma(x+1)$  for any  $x > 0$ . Condition (i) is verified.

**Case (B)**  $\beta'_0 := G'(S^c) < c_2\epsilon^r$ . Let  $\beta'_i = G'(B_i)$  for  $i = 1, \dots, k$ . Consider the map  $\Phi : \mathcal{P}(\Theta) \rightarrow \Delta^{k-1}$ , defined by

$$\Phi(Q) := (Q(B_1)/Q(S), \dots, Q(B_k)/Q(S)).$$

Define  $P_1 := \text{Dir}(\alpha\beta_1, \dots, \alpha\beta_k)$  and  $P_2 := \text{Dir}(\alpha'\beta'_1, \dots, \alpha'\beta'_k)$ . By a standard property of Dirichlet measures,  $P_1$  and  $P_2$  are push-forward measures of  $\mathcal{D}$  and  $\mathcal{D}'$ , respectively, by  $\Phi$ . (That is, if  $\text{law}(Q) = \mathcal{D}$ , then  $\text{law}(\Phi(Q)) = P_1$ . If  $\text{law}(Q) = \mathcal{D}'$  then  $\text{law}(\Phi(Q)) = P_2$ .) Define

$$B_1 := \left\{ \mathbf{q} \in \Delta^{k-1} \left| \frac{dP_2}{dP_1}(\mathbf{q}) > 1 \right. \right\}.$$

(This is exactly the same set defined by Eq. (22) in the proof of Lemma 4.1 that we shall encounter in the sequel). Now let  $\mathcal{B} = \Phi^{-1}(B_1)$ . Then, we have  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) = P_2(B_1) - P_1(B_1) = V(P_1, P_2)$ .

To verify condition (ii) of Definition 4.1, recall that

$$\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) = \mathcal{D} \left( \left\{ Q = \sum_{i=1}^k q_i \delta_{\theta_i} \left| Q \notin \mathcal{B} ; W_r(Q, Q') \leq \delta \text{ for some } Q' \in \mathcal{B} \right. \right\} \right).$$

For a measure of the form  $Q = \sum_{i=1}^k q_i \delta_{\theta_i}$ ,  $W_r(Q, Q') \leq \delta$  entails  $Q(B_i) - Q'(B_i) = q_i - Q'(B_i) \leq \delta^r / (c_1\epsilon)^r$ , and  $Q'(B_i) - q_i \leq \delta^r / (m - c_1\epsilon)^r < \delta^r / (c_1\epsilon)^r$ , for any  $i = 1, \dots, k$ . As well,  $Q'(S^c) \leq \delta^r / (c_1\epsilon)^r$ . This implies that,

$$|Q(B_i)/Q(S) - Q'(B_i)/Q'(S)| = \left| q_i - \frac{Q'(B_i)}{1 - Q'(S^c)} \right| \leq \frac{2\delta^r / (c_1\epsilon)^r}{1 - \delta^r / (c_1\epsilon)^r} \leq 4\delta^r / (c_1\epsilon)^r,$$

where the last inequality holds as soon as  $\delta \leq c_1\epsilon / 2^{1/r}$ . In short,  $W_r(Q, Q') \leq \delta$  implies that  $\|\Phi(Q) - \Phi(Q')\|_\infty \leq 4\delta^r / (c_1\epsilon)^r$ . We have

$$\begin{aligned} \mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) &\leq \mathcal{D} \left( \left\{ Q \left| Q \notin \mathcal{B}; \|\Phi(Q) - \Phi(Q')\|_\infty \leq 4\delta^r / (c_1\epsilon)^r \text{ for some } Q' \in \mathcal{B} \right. \right\} \right) \\ &= P_1(\{ \mathbf{q} \mid \mathbf{q} \notin B_1; \|\mathbf{q} - \mathbf{q}'\|_\infty \leq 4\delta^r / (c_1\epsilon)^r \text{ for some } \mathbf{q}' \in B_1 \}) \\ &\leq C_0(\delta/\epsilon)^{\alpha^* r}. \end{aligned}$$

The equality in the previous display is due to the definition of  $\mathcal{B}$ , while the last inequality is essentially the proof of Lemma 4.1 (b).  $C_0$  is a positive constant dependent only on  $\mathcal{D}$ .

It remains to verify condition (i) in Definition 4.1. We have

$$\begin{aligned}
V(P_1, P_2) &= V(\mathcal{D}_{\sum_{i=1}^k \alpha \beta_i \delta_{\theta_i}}, \mathcal{D}_{\sum_{i=1}^k \alpha' \beta'_i \delta_{\theta_i}}) \\
&\geq \frac{1}{(2 \operatorname{diam}(\Theta))^r} W_r^r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\sum_{i=1}^k \alpha' \beta'_i \delta_{\theta_i}}) \\
&\geq \frac{1}{(2 \operatorname{diam}(\Theta))^r} W_r^r(G, \sum_{i=1}^k \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}). \tag{21}
\end{aligned}$$

The first inequality in the above display is due to Theorem 6.15 of Villani [2008] (cf. Eq. (23)), while the second inequality is due to Lemma 3.1 (a). Now, we have

$$\begin{aligned}
W_r^r(G', \sum_{i=1}^k \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}) &\leq (c_1 \epsilon)^r \sum_{i=1}^k (\beta'_i \wedge \frac{\beta'_i}{1 - \beta'_0}) + \operatorname{diam}(\Theta)^r \sum_{i=1}^k \left| \beta'_i - \frac{\beta'_i}{1 - \beta'_0} \right| \\
&\leq (c_1 \epsilon)^r + \operatorname{diam}(\Theta)^r \sum_{i=1}^k \frac{\beta'_i \beta'_0}{1 - \beta'_0} \\
&\leq \epsilon^r (c_1^r + c_2 \operatorname{diam}(\Theta)^r) \leq \epsilon^r / 2^r.
\end{aligned}$$

The last inequalities in the above display is due to the hypothesis that  $\beta'_0 < c_2 \epsilon^r$ , and the choice of  $c_1, c_2$ . By triangle inequality,

$$W_r(G, \sum_{i=1}^k \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}) \geq W_r(G, G') - W_r(G', \sum_{i=1}^k \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}) \geq \epsilon - \epsilon/2 = \epsilon/2.$$

Combining with Eq. (21), we obtain that  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) = V(P_1, P_2) \geq \frac{1}{(2 \operatorname{diam}(\Theta))^r} (\epsilon/2)^r$ . This concludes the proof.  $\square$

The following lemma, which establishes strong regularity for a restricted class of Dirichlet measures, supplies a key argument in the proof of the previous theorem.

**Lemma 4.1.** *Let  $\mathcal{D} = \mathcal{D}_{\alpha G}$ , where  $G = \sum_{i=1}^k \beta_i \delta_{\theta_i}$  for some  $k < \infty$ ,  $\alpha, \alpha' > 0$ . Define*

$$\mathcal{C} = \{\mathcal{D}_{\alpha' G'} | G' \in \mathcal{P}(\Theta), \operatorname{spt} G' = \operatorname{spt} G\}.$$

- (a) *If  $\min_i \alpha \beta_i \geq 1$ , then  $\mathcal{C}$  has strong  $r$ -regular boundary with respect to  $\mathcal{D}$ .*
- (b) *If  $\max_i \alpha \beta_i < 1$ , then  $\mathcal{C}$  has strong  $\alpha^* r$ -regular boundary with respect to  $\mathcal{D}$ , where  $\alpha^* = \min_i \alpha \beta_i$ .*

*Proof.* (a) A random measure  $Q$  distributed by  $\mathcal{D}$  can be represented (with probability one) as  $Q = \sum_{i=1}^k q_i \delta_{\theta_i}$ , where  $\mathbf{q} = (q_1, \dots, q_k)$  is a  $k$ -dimensional Dirichlet vector:  $\operatorname{law}(\mathbf{q}) = \operatorname{Dir}(\alpha \beta_1, \dots, \alpha \beta_k)$ . Both  $\mathcal{D}$  and  $\mathcal{D}'$  are supported by the set of probability measures which are supported by  $\theta_1, \dots, \theta_k$ . In fact,  $\mathcal{D}$  and  $\mathcal{D}'$  are absolute continuous with each other and admit Radon-Nikodym density ratio derived from those of the Dirichlet distributions

$\text{Dir}(\alpha\beta_1, \dots, \alpha\beta_k)$  and  $\text{Dir}(\alpha'\beta'_1, \dots, \alpha'\beta'_k)$ . Let  $\mathcal{B}$  be the set of measures at which the density value under  $\mathcal{D}'$  is strictly greater than the density value under  $\mathcal{D}$ . The boundary set  $\text{bd } \mathcal{B}$  consists of those whose densities are equal under  $\mathcal{D}$  and  $\mathcal{D}'$ : let  $\Delta_i := \alpha\beta_i - \alpha'\beta'_i$ , then

$$\text{bd } \mathcal{B} = \left\{ \sum_{i=1}^k q_i \delta_{\theta_i} \left| \sum_{i=1}^k \Delta_i \log q_i = \log \frac{\Gamma(\alpha)}{\Gamma(\alpha')} + \sum_{i=1}^k \log \Gamma(\alpha'\beta'_i) - \log \Gamma(\alpha\beta_i) \right. \right\}. \quad (22)$$

We have  $V(\mathcal{D}', \mathcal{D}) = \mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) = V(\text{Dir}(\alpha\beta_1, \dots, \alpha\beta_k), \text{Dir}(\alpha'\beta'_1, \dots, \alpha'\beta'_k))$ . Since the Wasserstein distance is bounded by the weighted total variation, (cf. Theorem 6.15 of Villani [2008]),

$$W_r^r(\mathcal{D}, \mathcal{D}') \leq 2^{r-1} \int W_r^r(G_0, G) |\mathcal{D} - \mathcal{D}'|(dG) \leq 2^r \text{diam}(\Theta)^r V(\mathcal{D}, \mathcal{D}'). \quad (23)$$

Thus,  $V(\mathcal{D}, \mathcal{D}') \geq W_r^r(\mathcal{D}, \mathcal{D}') / (2 \text{diam}(\Theta))^r = W_r^r(G, G') / (2 \text{diam}(\Theta))^r$ , where the second equality is due to Lemma 3.1. So, condition (i) in Definition 4.1 is satisfied.

Condition (iii) will be established by deriving an upper bound for the measure of the  $\epsilon$ -tube  $\mathcal{D}((\text{bd } \mathcal{B})_\epsilon)$  (as a immediate consequence of Lemma 4.2, which we state following the end of this proof). Our proof strategy is to construct a suitable  $\epsilon$ -covering in  $\ell_\infty$  norm for  $(\text{bd } \mathcal{B})_\epsilon$ , and then establish upper bounds for the measure of each of the associated radius- $\epsilon$   $\ell_\infty$  balls.

Let  $\mathcal{Q}$  be the set of vectors  $\mathbf{q} = (q_1, \dots, q_k) \in \Delta^{k-1}$  that defines  $\text{bd } \mathcal{B}$  by Eq. (22).  $\mathcal{Q}$  can be partitioned into disjoint subsets, according to smallest and largest elements among  $\Delta_i/q_i$  for  $i = 1, \dots, k$ . There are  $k(k-1)$  such subsets, some of which may be empty. Consider one such subset, say  $S \subset \mathcal{Q}$  for which  $\Delta_1/q_1$  is the largest and  $\Delta_2/q_2$  is the smallest. Suppose that within set  $S$ ,  $\Delta_1/q_1 > \Delta_2/q_2$  (strictly greater than). By the implicit function theorem,  $q_1, q_2$  can be written as differentiable functions of  $q_3, \dots, q_k$  such that for each  $i = 3, \dots, k$ :

$$\begin{aligned} \frac{\partial q_1}{\partial q_i} &= -\frac{\Delta_i/q_i - \Delta_2/q_2}{\Delta_1/q_1 - \Delta_2/q_2}, \\ \frac{\partial q_2}{\partial q_i} &= \frac{\Delta_i/q_i - \Delta_1/q_1}{\Delta_1/q_1 - \Delta_2/q_2}. \end{aligned}$$

This shows that both  $|\partial q_1 / \partial q_i| \leq 1$  and  $|\partial q_2 / \partial q_i| \leq 1$  for all  $\mathbf{q} \in S$ . It follows that the number of radius- $\epsilon$   $\ell_\infty$  balls in  $\Delta^{k-1}$  needed to cover  $S$  is less than  $(1/\epsilon)^{k-2}$ . If, however,  $\Delta_1/q_1 = \Delta_i/q_i$  for all  $i = 2, \dots, k$ , then set  $S$  is a singleton, which is trivially covered by a single  $\ell_\infty$  ball. Since  $\mathcal{Q}$  is partitioned into at most  $k(k-1)$  subsets such as  $S$ , the  $\epsilon$ -covering number in  $\ell_\infty$  metric for  $\mathcal{Q}$  is less than  $k^2(1/\epsilon)^{k-2}$ .

It is simple to see that for any  $Q$  in the  $\epsilon$ -tube set

$$(\text{bd } \mathcal{B})_\epsilon \cap \text{spt } \mathcal{D} = \left\{ Q = \sum_{i=1}^k q_i \delta_{\theta_i} \left| W_r(Q, Q') \leq \epsilon \text{ for some } Q' \in \text{bd } \mathcal{B} \right. \right\}$$



there must be  $Q' = \sum_{i=1}^k q'_i \delta_{\theta_i} \in \text{bd } \mathcal{B}$  such that the  $\ell_\infty$  distance  $\|\mathbf{q} - \mathbf{q}'\|_\infty \leq \epsilon^r/m^r$ , where  $m = \min_{i,j \leq k} \|\theta_i - \theta_j\|$ . Let  $\mathcal{Q}^*$  be a  $\delta$ -covering of  $\mathcal{Q}$  in  $\ell_\infty$  metric, where  $\delta = \epsilon^r/m^r$ . For any  $Q \in (\text{bd } \mathcal{B})_\epsilon \cap \text{spt } \mathcal{D}$ , by triangle inequality there is a  $\mathbf{q}^* \in \mathcal{Q}^*$  such that  $\|\mathbf{q} - \mathbf{q}^*\|_\infty \leq 2\delta$ . Thus, let  $P_1$  denote the probability measure  $\text{Dir}(\alpha\beta_1, \dots, \alpha\beta_k)$  on the  $(k-1)$  dimensional simplex, then

$$\mathcal{D}((\text{bd } \mathcal{B})_\epsilon) \leq \sum_{\mathbf{q}^* \in \mathcal{Q}^*} P_1(\{\mathbf{q} : \|\mathbf{q} - \mathbf{q}^*\|_\infty \leq 2\delta\}). \quad (24)$$

Given that  $\alpha\beta_i \geq 1$  for all  $i = 1, \dots, k$ ,  $P_1(\{\mathbf{q} : \|\mathbf{q} - \mathbf{q}^*\|_\infty \leq 2\delta\}) \leq C(2\delta)^{k-1}$  where  $C$  is a universal constant. As a result,  $\mathcal{D}((\text{bd } \mathcal{B})_\epsilon) \leq k^2(1/\delta)^{k-2} \times C(2\delta)^{k-1} = C_0 \epsilon^r$  for  $C_0 = k^2 2^{k-1} C/m^r$ .

(b) This part requires a more refined estimate of the upper bound for  $\mathcal{D}$  measure on the  $\epsilon$ -tube set  $(\text{bd } \mathcal{B})_\epsilon$ . As in part (a),  $\mathcal{Q}^*$  denotes a  $\delta$ -covering of  $\mathcal{Q}$  in  $\ell_\infty$  metric. Consider an element  $\mathbf{q}^* \in \mathcal{Q}^*$ . There is one index  $j^* = j^*(\mathbf{q}^*) \in [1, k]$  such that  $q_{j^*}^* \geq 1/k$ . Then for  $\delta < 1/2k$ ,  $|q_{j^*}^* - q_{j^*}^*| \leq \delta$  entails  $q_{j^*}^* \geq 1/k - \delta > 1/2k$ . Now, apply the fact that  $0 \geq u \geq v$  and  $1 \geq a \geq b$  yields  $a^u \leq a^v \leq b^v$ , and  $\alpha^* - 1 \leq \alpha\beta_{j^*} - 1 \leq 0$  to obtain  $q_{j^*}^{\alpha\beta_{j^*}-1} \leq (1/2k)^{\alpha^*-1}$ . Under  $\text{Dir}(\alpha\beta_1, \dots, \alpha\beta_k)$ , the measure on the  $\ell_\infty$  ball of radius  $\delta$  and that centers at  $\mathbf{q}^*$  is:

$$\begin{aligned} & \mathbb{P}(\|\mathbf{q} - \mathbf{q}^*\|_\infty \leq \delta) \\ &= \frac{\Gamma(\alpha)}{\prod_{i=1}^k \Gamma(\alpha\beta_i)} \int_{\mathbf{q} \in \Delta^{k-1}; \|\mathbf{q} - \mathbf{q}^*\|_\infty \leq \delta} q_1^{\alpha\beta_1-1} \dots q_k^{\alpha\beta_k-1} dq_1 \dots dq_{k-1} \\ &\leq \frac{(1/2k)^{\alpha^*-1} \Gamma(\alpha)}{\prod_{i=1}^k \Gamma(\alpha\beta_i)} \prod_{i \neq j^*} \int_{q_i \in [(q_i^* - \delta)_+, (q_i^* + \delta)_{++}]} q_i^{\alpha\beta_i-1} dq_i \\ &= C(\alpha, \beta, k) \prod_{i \neq j^*} \left[ (q_i^* + \delta)_{++}^{\alpha\beta_i} - (q_i^* - \delta)_+^{\alpha\beta_i} \right] / (\alpha\beta_i). \end{aligned}$$

In the above display,  $(a)_+ := a \vee 0$ ,  $(a)_{++} := a \wedge 1$ , and  $C(\alpha, \beta, k) := \frac{(1/2k)^{\alpha^*-1} \Gamma(\alpha)}{\prod_{i=1}^k \Gamma(\alpha\beta_i)}$ .

Define  $A(q_i^*, \delta) := (3\delta)^{\alpha\beta_i} / \alpha^*$  if  $q_i^* \leq 2\delta$ , and  $A(q_i^*, \delta) := 2\delta[(t-1)\delta]^{\alpha\beta_i-1} = 2(t-1)^{\alpha\beta_i-1} \delta^{\alpha\beta_i}$  if  $q_i^* \in (t\delta, (t+1)\delta]$  for some natural number  $t \geq 2$ . It is simple to verify that, since  $\alpha\beta_i \leq 1$ ,

$$[(q_i^* + \delta)_{++}^{\alpha\beta_i} - (q_i^* - \delta)_+^{\alpha\beta_i}] / (\alpha\beta_i) \leq A(q_i^*, \delta).$$

So, we obtained an upper bound for the quantity in the previous display:

$$\mathbb{P}(\|\mathbf{q} - \mathbf{q}^*\|_\infty \leq \delta) \leq C(\alpha, \beta, k) \prod_{i \neq j^*} A(q_i^*, \delta).$$

Combining the above display with Eq. (24), for any  $\delta$ -covering  $\mathcal{Q}^*$  for  $\mathcal{Q}$ ,

$$\mathcal{D}((\text{bd } \mathcal{B})_\epsilon) \leq C(\alpha, \beta, k) \sum_{\mathbf{q}^* \in \mathcal{Q}^*} \prod_{i \neq j^*} A(q_i^*, 2\delta). \quad (25)$$

Let us consider a specific covering  $\mathcal{Q}^* = \{\mathbf{q}^* = (q_1^*, \dots, q_k^*)\} \subset \mathcal{Q}$  where  $q_i^*$ 's take values in the set  $\{t\delta | t \in \mathbb{N}, t \leq 1/\delta\}$ . Recall from part (a) that  $\mathcal{Q}$  can be partitioned into subsets  $S$ , each of which belongs to a  $k-2$  dimensional manifold such that two elements among the  $q_i^*$  for  $i = 1, \dots, k$  can be uniquely determined by the remaining  $k-2$  elements. Let  $i^* \neq j^*$  be the index of one of those two elements.

By the definition of  $A$ , if  $q_{i^*}^* \in (t\delta, (t+1)\delta]$  for some  $t \geq 2$ ,  $A(q_{i^*}^*, \delta) \leq 2\delta^{\alpha\beta_{i^*}} \leq 2\delta^{\alpha^*}$ . If  $q_{i^*}^* \leq 2\delta$ ,  $A(q_{i^*}^*, \delta) \leq (3\delta)^{\alpha^*}/\alpha^*$  for  $\delta < 1/3$ . So in any case,  $A(q_{i^*}^*, \delta) \leq 2(3\delta)^{\alpha^*}/\alpha^*$  as soon as  $\delta < 1/3$ . Hence, for  $\delta < 1/6$ ,

$$\sum_{\mathbf{q}^* \in \mathcal{Q}^*} \prod_{i \neq j^*} A(q_i^*, 2\delta) \leq 2(6\delta)^{\alpha^*}/\alpha^* \sum_{\mathbf{q}^* \in \mathcal{Q}^*} \prod_{i \neq j^*, i \neq i^*} A(q_i^*, 2\delta). \quad (26)$$

Subdivide set  $\mathcal{Q}^*$  into disjoint subsets according to the corresponding  $(i^*, j^*)$  pairs. There are at most  $k(k-1)/2$  such subsets. Consider one such subset, say  $\mathcal{Q}_{12}^*$ , which is distinguished by  $(i^*, j^*) = (1, 2)$ . We have

$$\sum_{\mathbf{q}^* \in \mathcal{Q}_{12}^*} \prod_{i \neq 1, i \neq 2} A(q_i^*, 2\delta) \leq \prod_{i \neq 1, i \neq 2} \sum_{q_i^*} A(q_i^*, 2\delta).$$

For any  $i = 1, \dots, k$ , we have

$$\begin{aligned} \sum_{q_i^*} A(q_i^*, 2\delta) &\leq 2(6\delta)^{\alpha^*}/\alpha^* + 2(2\delta)^{\alpha\beta_i} \sum_{t=2}^{\lfloor 1/\delta \rfloor} (t-1)^{\alpha\beta_i-1} \\ &\leq 2(6\delta)^{\alpha^*}/\alpha^* + 2(2\delta)^{\alpha\beta_i} \int_1^{1/\delta} x^{\alpha\beta_i-1} dx \\ &= 2(6\delta)^{\alpha^*}/\alpha^* + 2^{\alpha\beta_i+1}/\alpha\beta_i \leq 6/\alpha^* \end{aligned}$$

for  $\delta < 1/6$ . It follows from Eq. (26) that

$$\begin{aligned} \sum_{\mathbf{q}^* \in \mathcal{Q}^*} \prod_{i \neq j^*} A(q_i^*, 2\delta) &\leq 2(6\delta)^{\alpha^*}/\alpha^* \sum_{i^*, j^*} \left\{ \sum_{\mathbf{q}^* \in \mathcal{Q}_{i^*j^*}^*} \prod_{i \neq i^*, i \neq j^*} A(q_i^*, 2\delta) \right\} \\ &\leq (6\delta)^{\alpha^*}/\alpha^* \times k(k-1)(6/\alpha^*)^{k-2}. \end{aligned}$$

Combining the above display with Eq. (25), we conclude that for  $\delta = \epsilon^r/m^r < 1/6$ , there holds

$$\begin{aligned} \mathcal{D}((\text{bd } \mathcal{B})_\epsilon) &\leq \sum_{\mathbf{q}^* \in \mathcal{Q}^*} \mathbb{P}(\|\mathbf{q} - \mathbf{q}^*\|_\infty \leq 2\delta) \\ &\leq C(\alpha, \beta, k) (6\delta)^{\alpha^*}/\alpha^* \times k(k-1)(6/\alpha^*)^{k-2} \\ &= \frac{(1/2k)^{\alpha^*-1} \Gamma(\alpha)}{\prod_{i=1}^k \Gamma(\alpha\beta_i)} (6\delta)^{\alpha^*}/\alpha^* \times k(k-1)(6/\alpha^*)^{k-2}. \end{aligned}$$

□

**Lemma 4.2.** Suppose that subset of probability measures  $\mathcal{B} \subset \mathcal{P}(\Theta)$  is defined by  $\mathcal{B} = \{Q \in \mathcal{P}(\Theta) | F(Q) \leq 0\}$ , where  $F : \mathcal{P}(\Theta) \rightarrow \mathbb{R}$  is a continuous functional (i.e.,  $F(Q) \rightarrow F(Q_0)$  whenever  $W_r(Q, Q_0)$  tends to 0). Then,

(a)  $\text{bd } \mathcal{B} = \text{bd } \mathcal{B}^c = \{Q \in \mathcal{P}(\Theta) | F(Q) = 0\}.$

(b) For any  $\epsilon > 0$ ,  $\mathcal{B}_\epsilon \setminus \mathcal{B} \subset (\text{bd } \mathcal{B})_\epsilon$  and  $(\mathcal{B}^c)_\epsilon \setminus \mathcal{B}^c \subset (\text{bd } \mathcal{B})_\epsilon.$

*Proof.* (a) is immediate from the definition and the continuity of functional  $F$ . (b) is proved if we can show that if  $Q \in \mathcal{B}$  and  $Q' \in \mathcal{B}^c$  such that  $W_r(Q, Q') \leq \epsilon$ , then there exists  $P \in \text{bd } \mathcal{B}$  such that  $W_r(Q, P) \leq \epsilon$  and  $W_r(Q', P) \leq \epsilon$ . Indeed, consider collection of measures  $Q_t = tQ + (1-t)Q'$  for  $t \in [0, 1]$ . By the convexity of Wasserstein metric,  $W_r(Q_t, Q) \leq tW_r(Q, Q) + (1-t)W_r(Q', Q) \leq \epsilon$ . Note also that  $F(Q_t)$  is a continuous function of  $t$ . If either  $F(Q_0) = 0$  or  $F(Q_1) = 0$  (that is,  $Q \in \text{bd } \mathcal{B}$  or  $Q' \in \text{bd } \mathcal{B}$ , respectively), then we are done. Otherwise, we have  $F(Q_0) > 0$  and  $F(Q_1) < 0$ . So there exists  $t \in (0, 1)$  such that  $F(Q_t) = 0$ . It follows that  $Q_t \in \text{bd } \mathcal{B}$  and that  $W_r(Q, Q_t) \leq \epsilon$  and  $W_r(Q', Q_t) \leq \epsilon$ .  $\square$

## 4.2 The case of infinite and geometrically sparse support

In this subsection we study a class of base measures  $G$  that have infinite support points, but that remain amenable to our analysis of regular boundaries. In particular, we consider the class of sparse measures on  $\Theta$  (either ordinary sparse or supersparse) given by Definition 2.1.

**Theorem 4.2.** Assume that  $\mathcal{D} = \mathcal{D}_{\alpha G}$  for some  $\alpha \in (0, 1]$ .  $\text{spt } G$  is a  $(c_1, c_2, K)$ -sparse subset of a bounded space  $\Theta$  and that  $G$  is a sparse measure equipped with gauge function  $g$ . Let  $\alpha_1 \geq \alpha_0 > 0$ . Then, for any  $\mathcal{D}' \in \mathcal{C}$ , where

$$\mathcal{C} = \{\mathcal{D}' = \mathcal{D}_{\alpha' G'} | G' \in \mathcal{P}(\Theta), \alpha' \in [\alpha_0, \alpha_1]\}$$

there exists a measurable set  $\mathcal{B} \subset \mathcal{P}(\Theta)$  for which

(i)  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) \gtrsim W_r^r(G, G'),$

(ii) for any  $\delta \lesssim W_r(G, G'),$

$$\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) \lesssim 24^{K(c_0 W_r(G, G'))} \times \left( \frac{\delta}{W_r(G, G')} \right)^{\text{arg}(c_0 W_r(G, G'))}.$$

Here,  $c_0$  and the multiplying constants in  $\lesssim$  and  $\gtrsim$  depend only on  $\mathcal{D}$ .

The proof of this result is similar to Theorem 4.1 and deferred to Section 10.

## 5 Upper bounds for Wasserstein distances of base measures

The main purpose of this section is to obtain an upper bound of distance of Dirichlet base measures  $W_r(G, G')$  in terms of the variational distance of the marginal densities of observed data  $V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$ . In particular, we will establish inequalities of the form: for a fixed  $G \in \mathcal{P}(\Theta)$  and any  $G' \in \mathcal{P}(\Theta)$ ,

$$W_r^r(G, G') \lesssim V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) + A_n(G, G'), \quad (27)$$

where  $A_n(G, G')$  is a quantity that tends to 0 as  $n \rightarrow \infty$ . The rate at which  $A_n(G, G')$  tends to 0 depends on the sparse structure of  $G$ , and the smoothness of the kernel density  $f(x|\theta)$ . The full details are given in the statement of Theorem 5.1. It is worth contrasting this to the relatively easier inequalities in the opposite direction, given by Lemma 3.2:  $V(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq h(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \lesssim nW_{2r}^{2r}(G, G')$  holds generally for any pair of  $G, G'$ .

The proof of inequality (27) hinges on the existence of a suitable set  $\mathcal{B}_n \subset \mathcal{P}(\Theta)$  measurable with respect to (the sigma algebra induced by) the observed variables  $Y_{[n]}$ , which can then be used to distinguish  $G'$  from  $G$ , in the sense that

$$W_r^r(G, G') \lesssim P_{Y_{[n]}|G'}(\mathcal{B}_n) - P_{Y_{[n]}|G}(\mathcal{B}_n) + A_n(G, G').$$

In the previous section, we have already shown the existence of subset  $\mathcal{B} \subset \mathcal{P}(\Theta)$  for which

$$W_r^r(G, G') \lesssim \mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}).$$

To link up this result to the desired bound (27), the missing piece of the puzzle is the existence of a point estimate for the mixing measures on the basis of observed variables  $Y_{[n]}$ . In the following we shall establish the existence of such point estimators, which admits finite-sample probability bounds that may also be of independent interest.

**Finite-sample probability bounds for deconvolution problem.** Let  $\mathcal{Q}$  be a subset of  $\mathcal{P}(\Theta)$ , and  $\mathcal{F} = \{Q * f | Q \in \mathcal{Q}\}$ . Let  $\mathcal{Q}_k \subset \mathcal{P}(\Theta)$  be subset of measures with at most  $k$  support points.  $\mathcal{F}_k = \{Q * f | Q \in \mathcal{Q}_k\}$ . Given an iid  $n$ -vector  $Y_{[n]} = (Y_1, \dots, Y_n)$  according to the convolution mixture density  $Q_0 * f$  for some  $Q_0 \in \mathcal{Q}$ . Let  $\eta_n$  be a sequence of positive numbers converging to zero. Following Wong and Shen [1995] we consider an  $\eta_n$ -MLE (maximum likelihood estimator)  $\hat{f}_n \in \mathcal{F}$  such that

$$\frac{1}{n} \sum_{i=1}^n \log \hat{f}_n(Y_i) \geq \sup_{g \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \log g(Y_i) - \eta_n.$$

By our construction, there exists  $\hat{Q}_n \in \mathcal{Q}$  such that  $\hat{f}_n = \hat{Q}_n * f$ .

**Lemma 5.1.** *Suppose that Assumption A1 holds for some  $r \geq 1, C_1 > 0$ . Let  $\eta_n$  satisfy  $\eta_n \leq c_1 \epsilon_n^2$ ,  $\epsilon_n \rightarrow 0$  at a rate to be specified. Then the  $\eta_n$ -MLE satisfies the following bound under  $Q_0 * f$ -measure, for any  $Q_0 \in \mathcal{Q}$ :*

$$\mathbb{P}(h(\hat{f}_n, Q_0 * f) \geq \epsilon_n) \leq 5 \exp(-c_2 n \epsilon_n^2), \quad (28)$$

$$\mathbb{P}(W_2(\hat{Q}_n, Q_0) \geq \delta_n) \leq 5 \exp(-c_2 n \epsilon_n^2), \quad (29)$$

where  $c_1, c_2$  are some universal positive constants.  $\epsilon_n$  and  $\delta_n$  are given as follows:

(a)  $\epsilon_n = C_2(\log n/n)^{r/2d}$ , if  $d > 2r$ ;  $\epsilon_n = C_2(\log n/n)^{r/(d+2r)}$  if  $d < 2r$ , and  $\epsilon_n = (\log n)^{3/4}/n^{1/4}$  if  $d = 2r$ .

(b)  $\epsilon_n = C_2 n^{-1/2} \log n$ , if  $\mathcal{Q} = \mathcal{Q}_k$  and  $\mathcal{F} = \mathcal{F}_k$  for some  $k < \infty$ .

(c) If  $f$  is ordinary smooth with parameter  $\beta > 0$ , then  $\delta_n = C_3 \epsilon_n^{\frac{1}{2+\beta d'}}$  for any  $d' > d$ .

(d) If  $f$  is supersmooth with parameter  $\beta > 0$ , then  $\delta_n = C_3 [-\log \epsilon_n]^{-1/\beta}$ .

Here,  $C_2, C_3$  are different constants in each case.  $C_2$  depends only on  $d, r, \Theta$  and  $C_1$ , while  $C_3$  depends only on  $d, \beta, \Theta$  and  $C_2$ .

*Proof.* Part (a) is an application of Theorem 2 of Wong and Shen [1995], which is restated as follows: Suppose that  $\epsilon = \epsilon_n$  satisfies the following inequality:

$$\int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \left[ \log N(u/c_3, \mathcal{F}, h) \right]^{1/2} du \leq c_4 n^{1/2} \epsilon^2. \quad (30)$$

where  $c_3$  and  $c_4$  are certain universal constants (cf. Theorem 1 of Wong and Shen [1995]). Then, for some universal constants  $c_1, c_2 > 0$ , if  $\eta_n \leq c_1 \epsilon_n^2$ , the following probability bound holds under  $Q_0 * f$ -measure, for any  $Q_0 \in \mathcal{Q}$ ,

$$\mathbb{P}(h(\hat{f}_n, Q_0 * f) \geq \epsilon_n) \leq 5 \exp(-c_2 n \epsilon_n^2).$$

It remains to verify the entropy condition (30) given the rates specified in the statement of the present lemma. We shall make use of the following entropy bounds (cf. Lemma 4 of Nguyen [2013a]):

$$\log N(2\delta, \mathcal{Q}, W_r) \leq N(\delta, \Theta, \|\cdot\|) \log(e + e \operatorname{diam}(\Theta)^r / \delta^r), \quad (31)$$

$$\log(2\delta, \mathcal{Q}_k, W_r) \leq k(\log N(\delta, \Theta, \|\cdot\|) + \log(e + e \operatorname{diam}(\Theta)^r / \delta^r)). \quad (32)$$

By Assumption A2 and Lemma 3.2, we have  $h^2(Q * f, Q' * f) \leq C_1 W_{2r}^{2r}(Q, Q')$ . This implies that

$$N(u/c_3, \mathcal{F}, h) \leq N((u^2/c_3^2 C_1)^{1/2r}, \mathcal{Q}, W_{2r}).$$

Since  $\Theta \subset \mathbb{R}^d$ ,  $N(\delta, \Theta, \|\cdot\|) \leq (\text{diam}(\Theta)/\delta)^d$ . So, by (31),

$$\begin{aligned} & \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \left[ \log N((u^2/c_3^2 C_1)^{1/2r}, \mathcal{Q}, W_{2r}) \right]^{1/2} du \\ & \leq \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} \left[ N\left(\frac{u^{1/r}}{2c_3^{1/r} C_1^{1/2r}}, \Theta, \|\cdot\| \right) \log(e + e \text{diam}(\Theta)^{2r} 2^{2r} c_3^2 C_1/u^2) \right]^{1/2} du \\ & \leq \int_{\epsilon^2/2^8}^{\sqrt{2}\epsilon} (2 \text{diam}(\Theta))^{d/2} c_3^{d/2r} C_1^{d/4r} u^{-d/2r} [\log(e + e \text{diam}(\Theta)^{2r} 2^{2r} c_3^2 C_1/u^2)]^{1/2} du. \end{aligned}$$

For Eq. (30) to hold, it suffices to have the right side of the inequality in the above display bounded by  $c_4 n^{1/2} \epsilon^2$ . Indeed, this is straightforward to check for the rates given in part (a) of the lemma.

Part (b) of the lemma is proved in the same way, by invoking a tighter bound on the covering number via Eq. (32). Parts (c) and (d) are immediate consequences of part (a) and (b) by invoking Theorem 2 of Nguyen [2013a].  $\square$

We are ready to prove the key theorem of this section.

**Theorem 5.1.** *Suppose that  $\Theta$  is a bounded subset of  $\mathbb{R}^d$ , (A1) holds for some  $C_1 > 0$  and some  $r \in [1, 2]$ . Let  $\delta_n$  and  $\epsilon_n$  be vanishing sequences for which Eq. (29) holds. Fix  $G \in \mathcal{P}(\Theta)$  and  $\alpha \in (0, 1]$ , while  $G'$  varies in  $\mathcal{P}(\Theta)$ . Let  $\alpha^* = \alpha \inf_{\theta \in \text{spt } G} G(\{\theta\})$ . Then, there are positive constants  $c_0, c_1, C_0$  depending only on  $G$ , and  $c_2 > 0$  a universal constant, such that for any  $G' \in \mathcal{P}(\Theta)$ ,  $\alpha' \in [\alpha_1, \alpha_0]$  given and  $n$  sufficiently large so that  $\delta_n \lesssim W_r(G, G')$ , the following holds:*

$$c_0 W_r^r(G, G') \leq V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) + 10 \exp(-c_2 n \epsilon_n^2) + A_n(W_r(G, G')),$$

where  $A_n(W_r(G, G'))$  takes the form:

$$A_n(\omega) = \begin{cases} C_0(2\delta_n/\omega)^{\alpha^* r} & \text{if } G \text{ has finite support,} \\ C_0 24^{K(c_1 \omega)} (2\delta_n/\omega)^{\alpha r g(c_1 \omega)} & \text{if } G \text{ is } (\gamma_1, \gamma_2, K)\text{-sparse with gauge } g. \end{cases} \quad (33)$$

*Proof.* Suppose that  $G$  has finite support. By Theorem 4.1 (applied for  $W_r$ ) there are positive constants  $C_0, c_0$  independent of  $G'$  such that for some measurable set  $\mathcal{B} \subset \mathcal{P}(\Theta)$ , (i)  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) \geq c_0 W_r^r(G, G')$  and (ii)  $\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) \leq C_0(\delta/W_r(G, G'))^{\alpha^* r}$  for all  $\delta \lesssim W_r(G, G')$ .

Recall that  $\hat{Q}_n$  is a point estimate of  $Q$  defined earlier in this section. By the definition of variational distance, for any  $\delta > 0$

$$V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \geq \mathbb{P}(\hat{Q}_n \in \mathcal{B}_\delta | G') - \mathbb{P}(\hat{Q}_n \in \mathcal{B}_\delta | G).$$

Here,  $\mathbb{P}(\cdot|G)$  is taken to mean the probability of an event given that the observations are generated according to the Dirichlet base measure  $G$ . Set  $\mathcal{B}_\delta := \{Q \in \mathcal{P}(\Theta) | \text{there is } Q' \in \mathcal{B} \text{ such that } W_r(Q, Q') \leq \delta\}$ . We have

$$\begin{aligned} \mathbb{P}(\hat{Q}_n \in \mathcal{B}_\delta | G') &\geq \mathbb{P}(\hat{Q}_n \in \mathcal{B}_\delta, W_r(\hat{Q}_n, Q) < \delta | G') \\ &\geq \mathbb{P}(Q \in \mathcal{B}, W_r(\hat{Q}_n, Q) < \delta | G') \\ &\geq \mathcal{D}'(\mathcal{B}) - \mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta | G'). \end{aligned}$$

We also have

$$\begin{aligned} \mathbb{P}(\hat{Q}_n \in \mathcal{B}_\delta | G) &\leq \mathbb{P}(\hat{Q}_n \in \mathcal{B}_\delta, W_r(\hat{Q}_n, Q) < \delta | G) + \mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta | G) \\ &\leq \mathbb{P}(Q \in \mathcal{B}_{2\delta} | G) + \mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta | G) \\ &= \mathcal{D}(\mathcal{B}_{2\delta}) + \mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta | G). \end{aligned}$$

Hence,

$$\begin{aligned} V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) &\geq \mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}_{2\delta}) - 2 \sup_{Q \in \mathcal{Q}} \mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta) \\ &\geq (\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B})) - \mathcal{D}(\mathcal{B}_{2\delta} \setminus \mathcal{B}) - 2 \sup_{Q \in \mathcal{Q}} \mathbb{P}(W_r(\hat{Q}_n, Q) \geq \delta). \end{aligned}$$

Since  $r \in [1, 2]$ ,  $W_r(\hat{Q}_n, Q) \leq W_2(\hat{Q}_n, Q)$ . Choose  $\delta := \delta_n$  such that Eq. (29) holds. Then, as soon as  $2\delta_n \lesssim W_r(G, G')$ , for some multiplying constant depending only on  $G$ , we have

$$V(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'}) \geq c_0 W_r^r(G, G') - C_0 (2\delta_n / W_r(G, G'))^{\alpha^* r} - 10 \exp(-c_2 n \epsilon_n^2).$$

The case that  $G$  has infinite support proceeds in a similar way by invoking Thm 4.2.  $\square$

## 6 Proof of Theorem 2.1

We are now ready to prove Theorem 2.1, that the posterior distribution of the base measure  $G$  given the  $m \times n$  data set  $Y_{[n]}^{[m]}$  concentrates most its mass toward the true base measure  $G_0$ . The basic structure of the proof consists of fairly standard calculations, by applying the inequalities established in the previous sections.

We shall adopt the same strategy for establishing the convergence of mixing measures that arise in mixture models [Nguyen, 2013a]. To proceed we need several additional notions. Given  $G \in \mathcal{P}(\Theta)$ , define the Wasserstein ball centered at  $G$  as:  $B_{W_r}(G, \delta) := \{G' \in \mathcal{P}(\Theta) : W_r(G, G') \leq \delta\}$ . A useful notion is the Hellinger information of Wasserstein metric for a given set:

**Definition 6.1.** Fix  $G_0 \in \mathcal{P}(\Theta)$  and  $\mathcal{G} \subset \mathcal{P}(\Theta)$ . For a fixed  $n$ , the sample size of  $Y_{[n]}$ , define the Hellinger information of  $W_r$  metric for set  $\mathcal{G}$  as a real-valued function on the positive reals  $\Psi_{\mathcal{G}, n} : \mathbb{R}_+ \rightarrow \mathbb{R}$ :

$$\Psi_{\mathcal{G}, n}(\delta) := \inf_{G \in \mathcal{G}; W_r(G_0, G) \geq \delta/2} h^2(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}). \quad (34)$$

Recall that  $p_{Y_{[n]}|G}$  is the marginal density of  $n$ -vector  $Y_{[n]}$  obtained by integrating out the generic random measure  $Q$  via Eq. (10). We also define  $\Phi_{\mathcal{G},n} : \mathbb{R}_+ \rightarrow \mathbb{R}$  to be an arbitrary non-negative valued function such that for any  $\delta > 0$ ,

$$\sup_{G, G' \in \mathcal{G}; W_r(G, G') \leq \Phi_{\mathcal{G},n}(\delta)} h^2(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq \Psi_{\mathcal{G},n}(\delta)/4.$$

At a high level,  $\Psi$  and  $\Phi$  are introduced to relate the Hellinger distance on the marginal densities,  $h(P_{Y_{[n]}|G}, P_{Y_{[n]}|G'})$ , and the Wasserstein distance on the base measures,  $W_r(G, G')$ , in both directions. They play useful roles in specifying sufficient conditions for the posterior concentration. In both definitions of  $\Phi$  and  $\Psi$ , we suppress the dependence on (the fixed)  $G_0$  to simplify notations. Note that if  $G_0 \in \mathcal{G}$ , it follows from the definition that  $\Phi_{\mathcal{G},n}(\delta) < \delta/2$ .

Consider a general probability model given as follows

$$G \sim \Pi_G, Q_1, \dots, Q_m | G \sim \mathcal{D}_G$$

$$Y_{[n]}^i | Q_i \sim Q_i * f \text{ for } i = 1, \dots, m,$$

where  $\mathcal{D}_G \in \mathcal{P}^{(2)}(\Theta)$  is parameterized by  $G$ .

Let  $\mathcal{G}$  denote the support of the prior  $\Pi_G$ . Suppose that as  $n \rightarrow \infty$  and  $m \rightarrow \infty$  at a certain rate relative to  $n$ , there is a large constant  $C$ , a sequence of scalars  $\epsilon_{mn} \rightarrow 0$  defined in terms of  $m$  and  $n$ , and  $m\epsilon_{mn}^2 \rightarrow \infty$ , and a sequence of positive scalars  $M_m$  according to which the following hold:

$$\sup_{G_1 \in \mathcal{G}} \log D(\Phi_{\mathcal{G},n}(\epsilon), \mathcal{G} \cap B_{W_r}(G_1, \epsilon/2), W_r) + \quad (35)$$

$$\log D(\epsilon/2, \mathcal{G} \cap B_{W_r}(G_0, 2\epsilon) \setminus B_{W_r}(G_0, \epsilon), W_r) \leq m\epsilon_{mn}^2 \quad \forall \epsilon \geq \epsilon_{mn},$$

$$\Pi_G(B_K(G_0, \epsilon_{mn})) \geq \exp[-m\epsilon_{mn}^2 C], \quad (36)$$

$$\Psi_{\mathcal{G},n}(M_m \epsilon_{mn}) \geq 8\epsilon_{mn}^2 (C + 4), \quad (37)$$

$$\exp(2m\epsilon_{mn}^2) \sum_{j \geq M_m} \exp[-m\Psi_{\mathcal{G},n}(j\epsilon_{mn})/8] \rightarrow 0. \quad (38)$$

Then, by Theorem 4 of Nguyen [2013b] (or similarly, Theorem 4 of Nguyen [2013a])  $\Pi_G(G : W_r(G_0, G) \geq M_m \epsilon_{mn} | Y_{[n]}^{[m]}) \rightarrow 0$  in  $P_{Y_{[n]}|G_0}^m$ -probability.

Condition (35) controls the packing number  $D$  in  $W_r$  metric for several subsets of  $\mathcal{G}$ . A related notion is the covering number  $N(\epsilon, \mathcal{G}, W_r)$ . Condition (36) controls the thickness of the prior, where the thickness is specified in terms of the Kullback-Leibler neighborhood defined by Eq. (20). Conditions (37) and (38) control the behavior of the Hellinger information for  $\mathcal{G}$  defined earlier. It remains to verify these sufficient conditions in the context of the hierarchical Dirichlet process model, specifically for  $\mathcal{D}_G = \mathcal{D}_{\alpha G}$ , and  $\Pi_G = \mathcal{D}_{\gamma H}$ , by appealing to the results established in Sections 3, 4, and 5.



**Step 1** Since  $\Theta$  is bounded, Assumption (A1) holds for  $r \geq 1$  also implies that it holds for  $r = 1$ . By Lemma 3.2(a),  $h^2(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq nC_1W_1(G, G')$ . By the definition of  $\Phi_{\mathcal{G},n}(\delta)$  for  $W_1$ , it suffices to set  $\Phi_{\mathcal{G},n}(\delta) := \Psi_{\mathcal{G},n}(\delta)/(4nC_1)$ .

To verify condition expressed by Eq. (35), let us first derive a lower bound on  $\Phi_{\mathcal{G},n}$ , by invoking Theorem 5.1:

$$\begin{aligned}\Phi_{\mathcal{G},n}(\omega) &= \Psi_{\mathcal{G},n}(\omega)/(4nC_1) \geq \inf_{G:W_1(G_0,G) \geq \omega/2} V^2(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G})/(4nC_1) \\ &\geq \left[ c_0\omega/2 - A_n(\omega) - 10 \exp(-c_2n\epsilon_n^2) \right]^2 / (4nC_1),\end{aligned}$$

where the last inequality in the previous display holds whenever  $\delta_n \lesssim \omega$ .  $A_n$  is a decreasing function (in  $\omega$ ) given in (33). The quantities  $\epsilon_n$  and  $\delta_n$  are defined in Lemma 5.1, and constants  $c_0, c_1, c_2, C_0$  are dependent only on the fixed  $G_0$ . We shall check in the sequel that for the given rate  $\epsilon_{mn}$ , for any  $\omega > \epsilon_{mn}$ ,

$$A_n(\omega) + 10 \exp(-c_2n\epsilon_n^2) \leq c_0\epsilon_{mn}/4 < c_0\omega/4. \quad (39)$$

Suppose that (39) holds, then  $\Phi_{\mathcal{G},n}(\omega) \geq 2c\omega^2/n$  for some constant  $c > 0$  whenever  $\omega > \epsilon_{mn}$ . As a result,

$$\begin{aligned}\sup_{G_1 \in \mathcal{P}(\Theta)} \log D(\Phi_{\mathcal{G},n}(\omega), \mathcal{G} \cap B_{W_1}(G_1, \omega/2), W_1) &\leq \log N(c\omega^2/n, \mathcal{P}(\Theta), W_1) \\ &\leq (n \text{diam}(\Theta)/c\omega^2)^d \log(e + en \text{diam}(\Theta)/c\omega^2) \leq m\epsilon_{mn}^2/2,\end{aligned}$$

where the next-to-last equality in the previous display is an application (31), while the last inequality holds if  $\epsilon_{mn}$  satisfies:

$$\epsilon_{mn} \gtrsim [n^d \log(mn)/m]^{1/(2d+2)}.$$

With this condition on  $\epsilon_{mn}$ , for any  $\omega > \epsilon_{mn}$ ,

$$\begin{aligned}\log D(\omega/2, \mathcal{P}(\Theta), W_1) &\leq \log N(\omega/4, \mathcal{P}(\Theta), W_1) \\ &\leq N(\omega/8, \Theta, \|\cdot\|) \log(e + 8e \text{diam}(\Theta)/\omega) \\ &= (8 \text{diam}(\Theta)/\omega)^d \log(e + 8e \text{diam}(\Theta)/\omega) \leq m\epsilon_{mn}^2/2.\end{aligned}$$

Hence, Eq. (35) is verified.

**Step 2** Turning to condition (36), by Lemma 3.4,

$$\begin{aligned}&-\log \Pi(B_K(G_0, \epsilon_{mn})) \leq \\ &-D \log \gamma + [(1 + d/r)(D - 1) + Dd/r] \log(n^3/\epsilon_{mn}^2) + \log(1/c) \\ &\leq C(n^3/\epsilon_{mn}^2)^{d/r} \log(n^3/\epsilon_{mn}^2) \leq Cm\epsilon_{mn}^2\end{aligned}$$

for some constant  $C > 0$ , where the last inequality holds if

$$\epsilon_{mn} \gtrsim [n^{3d} \log(nm)/m]^{1/(2d+2)} \geq [n^{3d/r} \log(nm)/m]^{r/(2d+2r)}.$$

Since  $\Psi_{\mathcal{G},n}(\omega) = 4nC_1\Phi_{\mathcal{G},n}(\omega) \geq 8cC_1\omega^2$ , it is simple to check that both Eqs (37) and (38) will follow by setting  $M_m$  to be a sufficiently large constant.

It remains to clarify the condition expressed by Eq. (39).

**Step 3** First, consider the case  $G_0$  has finite support. Then, (39) is immediately satisfied if

$$\epsilon_{mn} \gtrsim \delta_n^{\alpha^*/(\alpha^*+1)} + \exp(-c_2 n \epsilon_n^2).$$

The second term in the right side of the previous display is relatively negligible compared to the first. So, when  $G_0$  has finite support, all sufficient conditions for the posterior concentration claim hold by setting

$$\epsilon_{mn} \gtrsim [n^d \log(mn)/m]^{1/(2d+2)} + \delta_n^{\alpha^*/(\alpha^*+1)}.$$

**Step 4** Next, consider the case  $G_0$  has infinite support, and in fact has geometrically sparse support. For the case that  $G_0$  is super sparse with parameters  $(\gamma_0, \gamma_1)$ , that is,  $K(\epsilon) \lesssim [\log(1/\epsilon)]^{\gamma_0}$ , and  $g(\epsilon) \gtrsim [\log(1/\epsilon)]^{-\gamma_1}$ , it is simple to verify that

$$A_n(\epsilon_{mn}) = 24^{K(c_1 \epsilon_{mn})} \times (2\delta_n/\epsilon_{mn})^{c_1 g(c_1 \epsilon_{mn})} \lesssim \epsilon_{mn}$$

if

$$\epsilon_{mn} > \exp - [\log(1/\delta_n)]^{1/(\gamma_1+1 \vee \gamma_0)}.$$

For the case that  $G_0$  is ordinary sparse with parameters  $(\gamma_0, \gamma_1)$ , that is  $K(\epsilon) \lesssim (1/\epsilon)^{\gamma_0}$ , and  $g(\epsilon) \gtrsim \epsilon^{\gamma_1}$ . It is simple to verify that  $A_n(\epsilon_{mn}) \lesssim \epsilon_{mn}$  if

$$\epsilon_{mn} > [\log(1/\delta_n)]^{-1/(\gamma_1+\gamma_0)}.$$

Examples of  $\epsilon_n$  and  $\delta_n$  are given in Lemma 5.1: If  $f$  is an ordinary smooth kernel density,  $\log(1/\delta_n) \asymp \frac{1}{2+\beta d'} \log(1/\epsilon_n) \asymp \log n$ . If  $f$  is a supersmooth kernel density,  $\log(1/\delta_n) \asymp \frac{1}{\beta} \log \log(1/\epsilon_n) \asymp \log \log n$ .

## 7 Borrowing strength in hierarchical Bayes

This section is devoted to the proof of Theorem 2.2. The proof is a simple consequence from Lemma 7.4, which establishes the posterior concentration behavior for a mixture distribution  $Q * f$ , where  $Q$  is a Dirichlet process distributed by  $\mathcal{D}_{\alpha G}$ , given that the base measure  $G$  is a small perturbation from the true base measure  $G_0$  that is now assumed to have finite support. A complete statement of Lemma 7.4 is given in section 7.3. In the following we proceed to give a proof of Theorem 2.2.

## 7.1 Proof of Theorem 2.2

Recall that for each  $\tilde{n}$ ,  $\epsilon_{mn} = \epsilon_{mn}(\tilde{n})$  is a net of scalars indexed by  $m, n$  that tend to 0. Define  $A_{mn}^{(\tilde{n})} := \{G : W_1(G, G_0) \geq \epsilon_{mn}\}$  and  $B_{mn}^{(\tilde{n})} := \{Q_0 : h(Q_0 * f, Q_0^* * f) \geq C((\log \tilde{n}/\tilde{n})^{1/(d+2)} + \epsilon_{mn}^{r/2} \log(1/\epsilon_{mn}))\}$  for some large constant  $C$ . Due to the conditional independence of  $Y_{[\tilde{n}]}^0$  and  $Y_{[n]}^{[m]}$  given  $G$ ,

$$\begin{aligned} \Pi_Q(Q_0 \in B_{mn}^{(\tilde{n})} | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}) &= \int \Pi_Q(Q_0 \in B_{mn}^{(\tilde{n})} | G, Y_{[\tilde{n}]}^0) d\Pi_G(G | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}) \\ &\leq \int_{\mathcal{P}(\Theta) \setminus A_{mn}^{(\tilde{n})}} \Pi_Q(Q_0 \in B_{mn}^{(\tilde{n})} | G, Y_{[\tilde{n}]}^0) d\Pi_G(G | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}) \\ &\quad + \Pi_G(G \in A_{mn}^{(\tilde{n})} | Y_{[\tilde{n}]}^0, Y_{[n]}^{[m]}). \end{aligned}$$

For each  $\tilde{n}$ , the second quantity in the upper bound tends to 0 in  $P_{Y_{[\tilde{n}]}^0 | Q_0^*} \times P_{Y_{[n]}^{[m]} | G_0}$ -probability, as  $m, n \rightarrow \infty$  at suitable rates by condition (b) of the theorem. Now, as  $\tilde{n} \rightarrow \infty$ , the first quantity tends to 0 as a consequence of Lemma 7.4. This completes the proof for (i). Parts (ii) and (iii) are proved in the same way.

## 7.2 Wasserstein geometry of the support of a single Dirichlet measure

Before proceeding to a proof for Lemma 7.4, we prepare three technical lemmas, which provide a detailed picture of the geometry of the support of a Dirichlet measure, and may be of independent interest. The first lemma demonstrates gains in the thickness of the conditional Dirichlet prior (given a perturbed base measure) compared to the unconditional Dirichlet prior. The second and third lemma show that Dirichlet measure concentrates most its mass on “small” sets, by which we mean sets that admit a small number of covering balls in Wasserstein metrics. This characterization enables the construction of a suitable sieves as required by the proof of Lemma 7.4.

**Lemma 7.1.** *Given  $G_0 = \sum_{i=1}^k \beta_i \delta_{\theta_i}$  and small  $\epsilon > 0$ . Let  $G \in \mathcal{P}(\Theta)$  such that  $W_1(G, G_0) \leq \epsilon$ . Suppose that  $\text{law}(Q) = \mathcal{D}_{\alpha G}$ , where  $\alpha \in (0, 1]$ .*

- (a) *For any  $Q_0 \in \mathcal{P}(\Theta)$  such that  $\text{spt } Q_0 \subset \text{spt } G_0$ , and any  $\delta$  such that  $\delta \geq \max_{i \leq k} 2\epsilon/\beta_i$  and  $\delta \leq \min_{i,j \leq k} \|\theta_i - \theta_j\|/2$ , any  $r \geq 1$ , there holds*

$$\mathbb{P}(W_r(Q_0, Q) \leq 2^{1/r} \delta) \geq \Gamma(\alpha)(\alpha/2)^k \left( \frac{\delta^r}{2k \text{diam}(\Theta)} \right)^{\alpha+k-1} \prod_{i=1}^k \beta_i.$$

- (b) *In addition, suppose that (A1-A2) hold for some  $r \geq 1$ . Then, there are constants  $C, c > 0$  depending only on  $\alpha, k, C_1, M, \text{diam}(\Theta), r$  and  $\beta_i$ 's such that for any  $\delta$  such that  $\delta/\log(1/\delta) \geq C\epsilon^{r/2}$ ,*

$$\mathbb{P}(Q \in B_K(Q_0, \delta)) \geq c(\delta/\log(1/\delta))^{2(\alpha+k-1)}.$$

This should be contrasted with the general small ball probability bound of Dirichlet process as stated by Lemma 3.3. In that lemma the base measure is an arbitrary non-atomic measure, while the lower bound is applied to any small  $W_r$  ball centering at an arbitrary measure. The lower bound is exponentially small in the radius. In the present lemma, the base measure  $G$  is constrained to being close to a discrete measure  $G_0$  with  $k < \infty$  support points, while the lower bound is applied to small  $W_r$  balls centering at  $Q_0$  that shares the same support as  $G_0$ . As a result, the lower bound is only polynomially small in the radius.

The following lemma relies on the intuition that the Dirichlet measure concentrates most its mass on probability measures which place most their mass on a “small” number of support points.

**Lemma 7.2.** *Let  $\mathcal{D} := \mathcal{D}_{\alpha G}$  and  $r \geq 1$ . For any  $\delta > 0$ , and for any  $k \in \mathbb{N}_+$ , there is a measurable set  $\mathcal{B}_k \subset \mathcal{P}(\Theta)$  satisfies the following properties:*

- (a)  $\sup_{Q \in \mathcal{B}_k} \inf_{Q' \in \mathcal{Q}_k} W_r(Q, Q') \leq \delta$ .
- (b)  $\log N(\delta, \mathcal{B}_k, W_r) \leq k(\log N(\delta/4, \Theta, \|\cdot\|) + \log(e + 4e \text{diam}(\Theta)^r / \delta^r))$ .
- (c) *There holds*

$$\mathcal{D}(\mathcal{P}(\Theta) \setminus \mathcal{B}_k) \leq k^{-k} (\delta / \text{diam}(\Theta))^{\alpha r} [e \alpha r \log(\text{diam}(\Theta) / \delta)]^k.$$

To see that the set  $\mathcal{B}_k$  has small entropy relative to  $\mathcal{P}(\Theta)$ , we note a general estimate for  $\mathcal{P}(\Theta)$ , which gives an upper bound that is exponentially large in terms of the entropy of  $\Theta$  (cf. Eq. (31)):

$$\log N(\delta, \mathcal{P}(\Theta), W_r) \leq N(\delta/2, \Theta, \|\cdot\|) \log(e + 2e \text{diam}(\Theta)^r / \delta^r).$$

In Lemma 7.2, the bound on entropy of  $\mathcal{B}_k$  increases only linearly in the entropy of  $\Theta$ . However, it also increases with  $k$ , which controls the measure of the complement of  $\mathcal{B}_k$ . Next, we consider the additional assumption that the Dirichlet base measure is a small perturbation of a discrete measure with  $k$  support points. The strength of this result compared to the previous lemma is that the entropy estimate depends only linearly on the entropy of  $\Theta$ , while  $k$  is fixed. The measure of the complement set of  $\mathcal{B}$  is controlled only by the amount of perturbation.

**Lemma 7.3.** *Given  $\epsilon > 0$ ,  $k < \infty$ ,  $r \geq 1$ . Let  $G_0, G \in \mathcal{P}(\Theta)$  such that that  $G_0$  has  $k$  support points and  $W_1(G, G_0) \leq \epsilon$ . Let  $\mathcal{D} := \mathcal{D}_{\alpha G}$  for some  $\alpha > 0$ . For any  $\delta > 0$ , there is a measurable set  $\mathcal{B} \subset \mathcal{P}(\Theta)$  that satisfies the following:*

- (a)  $\log N(\delta, \mathcal{B}, W_r) \leq k(\log N(\delta/4, \Theta, \|\cdot\|) + \log(e + 4e \text{diam}(\Theta)^r / \delta^r))$ .
- (b)  $\mathcal{D}(\mathcal{P}(\Theta) \setminus \mathcal{B}) \leq \epsilon \text{diam}(\Theta)^{r-1} / \delta^r$ .

The proofs of all three lemmas are given in Section 9.

### 7.3 Posterior concentration under perturbation of base measure

Here we state a key result that is needed in the proof of Theorem 2.2.

**Lemma 7.4.** *Let  $\Theta$  be a bounded subset of  $\mathbb{R}^d$ . Assumptions (A1-A2) hold. Let  $Q_0 \in \mathcal{P}(\Theta)$  such that  $\text{spt } Q_0 \subset \text{spt } G_0$ , where  $G_0 = \sum_{i=1}^k \beta_i \delta_{\theta_i}$  for some  $k < \infty$ . Let  $\Pi_G$  be an arbitrary prior distribution on  $\mathcal{P}(\Theta)$ . Consider the following hierarchical model*

$$\begin{aligned} G &\sim \Pi_G, \quad Q|G \sim \Pi_Q := \mathcal{D}_{\alpha G}, \\ Y_{[n]} = (Y_1, \dots, Y_n)|Q &\stackrel{iid}{\sim} Q * f. \end{aligned}$$

Let  $\epsilon_n \downarrow 0$  and define events  $\mathcal{E}_n := \{W_1(G, G_0) \leq \epsilon_n\}$ . Then, the posterior distribution of  $Q$  given  $Y_{[n]}$  admits the following as  $n \rightarrow \infty$ :

$$\Pi_Q \left( h(Q * f, Q_0 * f) \geq \delta_n \middle| Y_{[n]}, \mathcal{E}_n \right) \rightarrow 0, \quad (40)$$

$$\Pi_Q \left( W_2(Q, Q_0) \geq M_n \delta_n \middle| Y_{[n]}, \mathcal{E}_n \right) \rightarrow 0 \quad (41)$$

in  $(Q_0 * f) \times \Pi_G$ -probability, where the rates  $\delta_n$  and  $M_n \delta_n$  are given as follows:

- (i)  $\delta_n \asymp (\log n/n)^{1/(d+2)} + \epsilon_n^{r/2} \log(1/\epsilon_n)$ .
- (ii) If  $f$  is ordinary smooth with smoothness  $\beta > 0$ ,  $M_n \delta_n \asymp \delta_n^{\frac{1}{2+\beta d'}}$  for any  $d' > d$ .
- (iii) If  $f$  is supersmooth with smoothness  $\beta > 0$ , then  $M_n \delta_n \asymp (-\log \delta_n)^{-1/\beta}$ .

If  $\epsilon_n \downarrow 0$  suitably fast, then the following rates for  $\delta_n$  are valid:

- (iv) If  $f$  is ordinary smooth, and  $\epsilon_n \rightarrow 0$  sufficiently fast such that  $\epsilon_n \lesssim n^{-(\alpha+k+4M_0)} (\log n)^{-(\alpha+k-2)}$ , where  $M_0$  is some large constant, then  $\delta_n \asymp (\log n/n)^{1/2}$ .
- (v) If  $f$  is supersmooth with smoothness  $\beta > 0$ , and  $\epsilon_n \rightarrow 0$  sufficiently fast such that  $\epsilon_n \lesssim n^{-2(\alpha+k)/(\beta+2)} (\log n)^{-2(\alpha+k-1)} \exp(-4n^{\beta/(\beta+2)})$ , then  $\delta_n \asymp (1/n)^{1/(\beta+2)}$ .

We defer the proof of this lemma to Section 11. The basic structure contains of mostly standard calculations. The main novel part of the proof lies in the construction of suitable sieves that yield fast rates of convergence. The existence of such sieves is a direct consequence of the geometric lemmas presented in the previous subsection.

## 8 Proof of lemmas in Section 3

### 8.1 Proof of Lemma 3.1

(a) Take an arbitrary coupling  $\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')$  for which  $\int W_r^r(Q, Q') d\mathcal{K}$  is bounded. Let  $(Q, Q')$  be a pair of random measures whose law is  $\mathcal{K}$ . Given  $Q, Q'$ , let  $\kappa_{Q, Q'}$  denote an associated optimal coupling of  $(Q, Q')$  that is chosen in a measurable way (cf. Corollary 5.22 of Villani [2008]). By Fubini's theorem,

$$\begin{aligned} \int W_r^r(Q, Q') d\mathcal{K} &= \int \int \|\theta - \theta'\|^r d\kappa_{Q, Q'}(\theta, \theta') d\mathcal{K} \\ &= \int \|\theta - \theta'\|^r \int \kappa_{Q, Q'}(d\theta, d\theta') d\mathcal{K}. \end{aligned}$$

We note that the second integral in the last equation of the previous display can be written as  $\kappa(d\theta, d\theta')$ , for some valid coupling  $\kappa \in \mathcal{T}(G, G')$ . To see this, by marginalizing out  $\theta'$  we have for any measurable  $A \in \Theta$ ,

$$\int \kappa_{Q, Q'}(A \times \Theta) d\mathcal{K} = \int Q(A) d\mathcal{K} = \int Q(A) d\mathcal{D} = G(A).$$

The first equality is by the definition of  $\kappa_{Q, Q'}$ ; the second is by the definition of  $\mathcal{K}$ ; the third is by the assumption on  $\mathcal{D}$ . A similar identity holds for marginalizing out  $\theta$ . Thus,  $\int W_r^r(Q, Q') d\mathcal{K} \geq \inf_{\kappa \in \mathcal{T}(G, G')} \int \|\theta - \theta'\|^r d\kappa = W_r^r(G, G')$ . This inequality holds for any coupling  $\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')$ , so  $W_r^r(\mathcal{D}, \mathcal{D}') \geq W_r^r(G, G')$ .

(b) Let  $\kappa$  be an optimal coupling of  $(G, G')$ , i.e.,  $W_r^r(G, G') = \int \|\theta - \theta'\|^r d\kappa(\theta, \theta')$ , and  $\gamma$  be a random probability measure on  $\Theta \times \Theta$  such that  $\text{law}(\gamma) = \mathcal{D}_{\alpha\kappa}$ . Let  $Q, Q'$  be the random marginal measures induced by the joint measure  $\gamma(d\theta, d\theta')$  with respect to  $\theta$  and  $\theta'$ , respectively. By the definition of Dirichlet measures, along with the fact that  $\kappa \in \mathcal{T}(G, G')$ , we have  $\text{law}(Q) = \mathcal{D}_{\alpha G}$ , and  $\text{law}(Q') = \mathcal{D}_{\alpha G'}$ . Thus, the joint distribution of  $(Q, Q')$  is denoted by  $\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')$ . Since  $\gamma$  is a coupling of  $(Q, Q')$  by our construction, we have  $W_r^r(Q, Q') \leq \int \|\theta - \theta'\|^r d\gamma(\theta, \theta')$  almost surely. The expectation under the coupling  $\mathcal{K}$  satisfies:

$$\begin{aligned} \mathbb{E} W_r^r(Q, Q') &\leq \mathbb{E} \int \|\theta - \theta'\|^r d\gamma(\theta, \theta') \\ &= \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') = W_r^r(G, G'). \end{aligned}$$

The first equality is due to a standard property of Dirichlet measures that  $\mathbb{E}\gamma = \kappa$  and Fubini's theorem. We deduce that  $W_r^r(\mathcal{D}, \mathcal{D}') \leq W_r^r(G, G')$ . Combining with part (a) gives the desired identity.

### 8.2 Proof of Lemma 3.2

To simplify notations, let  $\mathcal{D} = \mathcal{D}_{\alpha G}$  and  $\mathcal{D}' = \mathcal{D}_{\alpha G'}$ . The density  $p_{Y_{[n]}|G}$ , defined by Eq. (10), is succinctly written as  $p_{Y_{[n]}|G} := \int (Q * f)^n \mathcal{D}(dQ)$ , and likewise,  $p_{Y_{[n]}|G'} :=$

$\int (Q' * f)^n \mathcal{D}'(dQ')$ . Due to the convexity of Kullback-Leibler divergence, by Jensen's inequality we obtain that for any coupling  $\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')$ ,

$$\begin{aligned} K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) &= K\left(\int (Q * f)^n \mathcal{K}(dQ, dQ'), \int (Q' * f)^n \mathcal{K}(dQ, dQ')\right) \\ &\leq \int K((Q * f)^n, (Q' * f)^n) \mathcal{K}(dQ, dQ') \\ &= n \int K(Q * f, Q' * f) \mathcal{K}(dQ, dQ'). \end{aligned}$$

Using the same argument, now for any coupling  $\kappa \in \tau(Q, Q')$

$$\begin{aligned} K(Q * f, Q' * f) &= K\left(\int f(\cdot|\theta) \kappa(d\theta, d\theta'), \int f(\cdot|\theta') \kappa(d\theta, d\theta')\right) \\ &\leq \int K(f(\cdot|\theta), f(\cdot|\theta')) \kappa(d\theta, d\theta') \\ &\leq \int C_1 \|\theta - \theta'\|^r \kappa(d\theta, d\theta'). \end{aligned}$$

Since this holds for any  $\kappa \in \tau(Q, Q')$ , we obtain that  $K(Q * f, Q' * f) \leq C_1 W_r^r(Q, Q')$ . Plugging back in the upper bound for  $K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'})$ , we have:

$$K(p_{Y_{[n]}|G}, p_{Y_{[n]}|G'}) \leq C_1 n \inf_{\mathcal{K} \in \mathcal{T}(\mathcal{D}, \mathcal{D}')} \int W_r^r(Q, Q') \mathcal{K}(dQ, dQ') = C_1 n W_r^r(\mathcal{D}, \mathcal{D}').$$

Up to this point we have not used the fact that  $\mathcal{D}$  and  $\mathcal{D}'$  are Dirichlet measures. By Lemma 3.1 (b) we arrive at the desired inequality. The proof of the other two inequalities are similar.

### 8.3 Proof of Lemma 3.4

We shall invoke a bound of Wong and Shen [1995] (Theorem 5) on the KL divergence. This bound says that if  $p$  and  $q$  are two densities on a common space such that  $\int p^2/q < M$ , then for some universal constant  $\epsilon_0 > 0$ , as long as  $h(p, q) \leq \epsilon < \epsilon_0$ , there holds:  $K(p, q) \leq C_0 \epsilon^2 \log(M/\epsilon)$ , and  $K_2(p, q) := \int p(\log(p/q))^2 \leq C_0 \epsilon^2 [\log(M/\epsilon)]^2$ , where  $C_0$  is a universal constant.

For a pair of  $G_0, G \in \mathcal{P}(\Theta)$ , if  $W_r(G_0, G) \leq \epsilon$  then by Lemma 3.2,

$$h^2(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq K(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G})/2 \leq C_1 n \epsilon^r / 2.$$

We also have  $\chi(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq M^n$ . We can apply the upper bound described in the previous paragraph to obtain:

$$K_2(p_{Y_{[n]}|G_0}, p_{Y_{[n]}|G}) \leq C_0 C_1 n \epsilon^r \left[ \frac{1}{2} \log \frac{2}{C_1 n \epsilon^r} + n \log M \right]^2.$$

If we set  $\epsilon^r = \delta^2/n^3$ , then the quantity in the right hand side of the previous display is bounded by  $C_2\delta^2$ , as long as  $n > \log(1/\delta) \vee (C_1\delta/2\epsilon_0)^{1/2}$ , where constant  $C_2$  depends only on  $C_0, C_1, M$ . Thus,

$$\mathbb{P}(G \in B_K(G_0, \delta)) \geq \mathbb{P}(W_r^r(G_0, G) \leq C_3\delta^2/n^3),$$

where constant  $C_3$  depends only on  $C_0, C_1, M$ . Combining with Lemma 3.3, and by (A3),  $H(S_i) \geq \eta_0\epsilon^d = \eta_0(\delta^2/n^3)^{d/r}$ , and  $D = (\text{diam}(\Theta)/\epsilon)^d$  to obtain the desired lower bound.

## 9 Proof of lemmas in subsection 7.2

### 9.1 Proof of Lemma 7.1

(a) Let  $B_i$  be the closed ball in  $\Theta$  with radius  $\delta > 0$  and centered at  $\theta_i$  for  $i = 1, \dots, k$ . Suppose that  $\delta$  is sufficiently small so that  $\delta < \min_{i,j \leq k} \|\theta_i - \theta_j\|/2$ . That is,  $B_i$ 's are disjoint subsets of  $\Theta$ . Let  $g_i = G(B_i)$  for  $i = 1, \dots, k$  and  $g_0 = 1 - \sum_{i=1}^k g_i$ . Since  $W_1(G, G_0) \geq \delta(\beta_i - g_i)$ , it follows that  $\beta_i - g_i \leq \epsilon/\delta$ . By the condition on  $\delta$ ,  $g_i \geq \beta_i - \epsilon/\delta \geq \beta_i/2$  for any  $i = 1, \dots, k$ . We also obtain  $g_0 \leq \epsilon/\delta$ .

Suppose that  $g_0 > 0$  (the case that  $g_0 = 0$  is handled similarly, yielding a stronger lower bound). Let  $q_i = Q(B_i)$  for  $i = 1, \dots, k$  and  $q_0 = 1 - \sum_{i=1}^k q_i$ . By assumption,  $\text{law}(\mathbf{q}) = \text{Dir}(\alpha \mathbf{g})$ . Let  $Q_0 = \sum_{i=1}^k q_i^* \delta_{\theta_i}$  and also set  $q_0^* := 0$ . If  $\|\mathbf{q} - \mathbf{q}^*\|_1 = \sum_{i=0}^k |q_i - q_i^*| < \delta^r / \text{diam}(\Theta)$ , then

$$W_r^r(Q, Q_0) \leq \delta^r \sum_{i=1}^k q_i \wedge q_i^* + \sum_{i=1}^k |q_i - q_i^*| \text{diam}(\Theta) \leq 2\delta^r.$$

Define  $\mathcal{E} = \{(q_1, \dots, q_k, q_0) \in \Delta^{k-1} \text{ such that } |q_i - q_i^*| \leq \frac{\delta^r}{2k \text{diam}(\Theta)} \text{ for } i = 0, \dots, k-1\}$ . If  $\mathbf{q} \in \mathcal{E}$  then  $\|\mathbf{q} - \mathbf{q}^*\|_1 \leq 2 \sum_{i=0}^{k-1} |q_i - q_i^*| \leq \delta^r / \text{diam}(\Theta)$ . It follows that

$$\begin{aligned} \mathbb{P}(W_r^r(Q, Q_0) \leq 2\delta^r) &\geq \mathbb{P}(\mathcal{E}) \\ &= \frac{\Gamma(\alpha)}{\prod_{i=0}^k \Gamma(\alpha g_i)} \int_{\mathcal{E}} q_0^{\alpha g_0 - 1} (1 - \sum_{i=0}^{k-1} q_i)^{\alpha g_k - 1} \prod_{i=1}^{k-1} q_i^{\alpha g_i - 1} dq_0 \dots dq_{k-1} \\ &\geq \frac{\Gamma(\alpha)}{\prod_{i=0}^k \Gamma(\alpha g_i)} \prod_{i=0}^{k-1} \int_{q_i \in ((q_i^* - \delta^r/2k \text{diam}(\Theta))_+, (q_i^* + \delta^r/2k \text{diam}(\Theta))_{++})} q_i^{\alpha g_i - 1} dq_i \\ &\geq \frac{\Gamma(\alpha)}{\prod_{i=0}^k \Gamma(\alpha g_i)} \times \frac{(\delta^r/2k \text{diam}(\Theta))^{\alpha g_0}}{\alpha g_0} \times (\delta^r/2k \text{diam}(\Theta))^{k-1}. \end{aligned}$$

Here,  $a_+ = a \vee 0$  and  $a_{++} = a \wedge 1$ ; the second and the third inequality in the above display are due to  $q_i^{\alpha g_i - 1} \geq 1$  for all  $i = 1, \dots, k$  as  $\alpha \leq 1$ . Finally, note that  $\Gamma(\alpha g_i)(\alpha g_i) =$



$\Gamma(\alpha g_i + 1) \leq 1$  for  $i = 0, \dots, k$ . So,

$$\begin{aligned} \mathbb{P}(W_r^r(Q, Q_0) \leq 2\delta^r) &\geq \Gamma(\alpha) \prod_{i=1}^k (\alpha g_i) \left( \frac{\delta^r}{2k \operatorname{diam}(\Theta)} \right)^{\alpha g_0 + k - 1} \\ &\geq \Gamma(\alpha) (\alpha/2)^k \left( \frac{\delta^r}{2k \operatorname{diam}(\Theta)} \right)^{\alpha + k - 1} \prod_{i=1}^k \beta_i. \end{aligned}$$

Part (b) follows from part (a). Its argument is similar to the proof of Lemma 3.4 is omitted.

## 9.2 Proof of Lemma 7.2

Thanks to Sethuraman [Sethuraman, 1994], a random measure whose law is  $\mathcal{D}$  may be parameterized as  $Q = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$ , where  $\theta_i$ 's are iid according to  $G$ , and  $p_i$ 's are distributed according to a “stick-breaking” process:  $p_1 = v_1, p_2 = v_2(1 - v_1), \dots, p_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$  for any  $k = 1, 2, \dots$ , where  $v_1, v_2, \dots$  are beta random variables iid according to  $\text{Beta}(1, \alpha)$ .

It is simple to check that  $1 - \sum_{i=1}^k p_i = \prod_{i=1}^k (1 - v_i)$ . By Markov's inequality, for any  $\epsilon > 0$ , under  $\mathcal{D}$  measure  $\mathbb{P}(1 - \sum_{i=1}^k p_i \geq \epsilon) = \mathbb{P}(\prod_{i=1}^k (1 - v_i) \geq \epsilon) \leq \inf_{t>0} \prod_{i=1}^k \mathbb{E}(1 - v_i)^t / \epsilon^t = \inf_{t>0} [\alpha / (\alpha + t)]^k / \epsilon^t = \exp[-\alpha \log(1/\epsilon) - k \log k + k \log(e\alpha) + k \log \log(1/\epsilon)] = \Delta(\epsilon, k)$ .

Define  $\mathcal{B}_k$  to be the subset of discrete measures  $Q$  that lie in the support of the Dirichlet measure, such that under stick-breaking representation described above,  $1 - \sum_{i=1}^k p_i < (\delta / \operatorname{diam}(\Theta))^r$ . Hence,  $\mathcal{D}(\mathcal{P}(\Theta) \setminus \mathcal{B}_k) \leq \Delta((\delta / \operatorname{diam}(\Theta))^r, k)$ . This concludes part (c).

For any  $Q = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$  such that  $1 - \sum_{i=1}^k p_i < \epsilon$ , by choosing  $Q' = \sum_{i=1}^k p_i \delta_{\theta_i} + (1 - \sum_{i=1}^k p_i) \delta_{\theta_k} \in \mathcal{G}_k(\Theta)$  we have  $W_r(Q, Q') \leq \epsilon^{1/r} \operatorname{diam}(\Theta)$ . Thus,  $\inf_{Q' \in \mathcal{Q}_k} W_r(Q, Q') \leq \epsilon^{1/r} \operatorname{diam}(\Theta) = \delta$ , concluding (a). For (b), we obtain that  $\log N(2\delta, \mathcal{B}_k, W_r) \leq \log N(\delta, \mathcal{Q}_k, W_r)$ . Combining with Eq. (32) to conclude.

## 9.3 Proof of Lemma 7.3

By Lemma 3.1(b),  $W_r^r(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G_0}) = W_r^r(G, G_0) \leq \epsilon \operatorname{diam}(\Theta)^{r-1}$ . This implies that there exists a coupling  $\mathcal{K} \in \mathcal{T}(\mathcal{D}_{\alpha G}, \mathcal{D}_{\alpha G_0})$  such that  $\int W_r^r(Q, Q') d\mathcal{K} \leq \epsilon \operatorname{diam}(\Theta)^{r-1}$ . Let  $\mathcal{Q}_0 \subset \mathcal{P}(\Theta)$  be the support of  $\mathcal{D}_{\alpha G_0}$  — this consists of probability measures that share the same  $k$  support points with  $G_0$ . Write  $W_r(Q, \mathcal{Q}_0) := \inf_{P \in \mathcal{Q}_0} W_r(Q, P)$ . Then  $\int W_r^r(Q, \mathcal{Q}_0) d\mathcal{D} = \int W_r^r(Q, \mathcal{Q}_0) d\mathcal{K} \leq \int W_r^r(Q, Q') d\mathcal{K} \leq \epsilon \operatorname{diam}(\Theta)^{r-1}$ . Let  $\mathcal{B} = \{Q : W_r(Q, \mathcal{Q}_0) < \delta\}$ . By Markov's inequality,  $\mathcal{D}(Q : W_r(Q, \mathcal{Q}_0) \geq \delta^r) \leq \epsilon \operatorname{diam}(\Theta)^{r-1} / \delta^r$ , yielding (b). Moreover,  $\log N(2\delta, \mathcal{B}, W_r) \leq \log(\delta, \mathcal{Q}_0, W_r)$ . Combining with Eq. (32) immediately yields (a).

## 10 Proof of Theorem 4.2

*Proof.* Take any  $G' \in \mathcal{P}(\Theta)$ . By definition, for any natural number  $t$ , there is  $\epsilon \in (c_1^{t+1}W_r(G, G'), c_1^tW_r(G, G'))$  such that  $\text{spt } G$  is covered by  $K := K(\epsilon)$  closed balls of radius  $\epsilon$ , to be denoted by  $B_1, \dots, B_K$ . In the following, let  $c_3 = c_2/2$ , and  $t$  be chosen such that  $c_1^t + ((1 + c_3)^r + \text{diam}(\Theta)^r)^{1/r} c_1^t < 1/2$ . Let  $B_i^* := (B_i)_{c_3\epsilon}$ , and  $S = \cup_{i \leq K} B_i^*$ . The proof proceeds in a similar fashion to that of Theorem 4.1. There are two scenarios: either (A)  $G'(S^c) \geq \epsilon^r$  or (B)  $G'(S^c) < \epsilon^r$ .

**Case (A).**  $\beta'_0 := G'(S^c) \geq \epsilon^r$ . Let  $\mathcal{B} = \{Q \in \mathcal{P}(\Theta) | Q(S^c) > 1/2\}$ , then  $\mathcal{D}(\mathcal{B}) = 0$ . Moreover, for any  $Q \in \mathcal{B}$  and  $Q' \in \text{spt } \mathcal{D}$ , so that  $\text{spt } Q' \subset S$ , we have  $W_r^r(Q, Q') \geq (1/2)(c_3\epsilon)^r$ . So, for any  $\delta < (1/2)^{1/r} c_3\epsilon$ ,  $\mathcal{D}(\mathcal{B}_\delta) = 0$ . This confirms (ii).

Suppose that  $\text{law}(Q) = \mathcal{D}'$ , then  $\text{law}(Q(S)) = \text{Beta}(\alpha'G'(S), \alpha'G'(S^c))$ . The same argument as that of Theorem 4.1 yields  $\mathcal{D}'(\mathcal{B}) \geq \frac{(1/2)^{2\alpha'} \Gamma(\alpha') \alpha' \epsilon^r}{\max_{1 \leq x \leq \alpha'+1} \Gamma(x)^2} \gtrsim W_r^r(G, G')$ , which confirms (i).

**Case (B).**  $\beta'_0 = G'(S^c) < \epsilon^r$ . Let  $\beta_i = G(B_i^*) = G(B_i)$  and  $\beta'_i = G'(B_i^*)$  for  $i = 1, \dots, K$  for  $K = K(\epsilon)$ . Consider the map  $\Phi : \mathcal{P}(\Theta) \rightarrow \Delta^{K-1}$ , defined by

$$\Phi(Q) := (Q(B_1^*)/Q(S), \dots, Q(B_K^*)/Q(S)).$$

Define  $P_1 := \text{Dir}(\alpha\beta_1, \dots, \alpha\beta_K)$  and  $P_2 := \text{Dir}(\alpha'\beta'_1, \dots, \alpha'\beta'_K)$ , and let

$$B_1 := \left\{ \mathbf{q} \in \Delta^{K-1} \left| \frac{dP_2}{dP_1}(\mathbf{q}) > 1 \right. \right\}.$$

Now define  $\mathcal{B} := \Phi^{-1}(B_1)$ . Then, we have  $\mathcal{D}'(\mathcal{B}) - \mathcal{D}(\mathcal{B}) = P_2(B_1) - P_1(B_1) = V(P_1, P_2)$ .

Let  $\theta_1, \dots, \theta_K$  be the center of the balls  $B_1, \dots, B_K$ , respectively. Then,

$$\begin{aligned} V(P_1, P_2) &= V(\mathcal{D}_{\sum_{i=1}^K \alpha\beta_i \delta_{\theta_i}}, \mathcal{D}_{\sum_{i=1}^K \alpha'\beta'_i \delta_{\theta_i}}) \\ &\geq \frac{1}{(2 \text{diam}(\Theta))^r} W_r^r(\mathcal{D}_{\sum_{i=1}^K \alpha\beta_i \delta_{\theta_i}}, \mathcal{D}_{\sum_{i=1}^K \alpha'\beta'_i \delta_{\theta_i}}) \\ &\geq \frac{1}{(2 \text{diam}(\Theta))^r} W_r^r\left(\sum_{i=1}^K \beta_i \delta_{\theta_i}, \sum_{i=1}^K \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}\right). \end{aligned} \quad (42)$$

The first inequality in the above display is due to Theorem 6.15 of Villani [2008] (cf. Eq. (23)), while the second inequality is due to Lemma 3.1 (a). Now, it is simple to see

that  $W_r(G, \sum_{i=1}^K \beta_i \delta_{\theta_i}) \leq \epsilon \leq c_1^t W_r(G, G')$ . Also,

$$\begin{aligned}
W_r^r(G', \sum_{i=1}^k \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}) &\leq (\epsilon + c_3 \epsilon)^r \sum_{i=1}^k (\beta'_i \wedge \frac{\beta'_i}{1 - \beta'_0}) + \text{diam}(\Theta)^r \sum_{i=1}^k \left| \beta'_i - \frac{\beta'_i}{1 - \beta'_0} \right| \\
&\leq (1 + c_3)^r \epsilon^r + \text{diam}(\Theta)^r \beta'_0 \\
&\leq ((1 + c_3)^r + \text{diam}(\Theta)^r) \epsilon^r \\
&\leq ((1 + c_3)^r + \text{diam}(\Theta)^r) c_1^{rt} W_r^r(G, G').
\end{aligned}$$

By triangle inequality,

$$\begin{aligned}
W_r(\sum_{i=1}^K \beta_i \delta_{\theta_i}, \sum_{i=1}^K \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}) &\geq W_r(G, G') - W_r(G, \sum_{i=1}^K \beta_i \delta_{\theta_i}) - W_r(G', \sum_{i=1}^K \frac{\beta'_i}{1 - \beta'_0} \delta_{\theta_i}) \\
&\geq \left[ 1 - c_1^t - ((1 + c_3)^r + \text{diam}(\Theta)^r)^{1/r} c_1^t \right] W_r(G, G') \geq W_r(G, G')/2.
\end{aligned}$$

Thus, claim (i) is established.

To prove (ii), note that for any pair  $Q, Q' \in \mathcal{P}(\Theta)$  such that  $Q \in \text{spt } \mathcal{D}$ ,  $W_r(Q, Q') \leq \delta$  entails  $Q'(B_i^*) - Q(B_i) \leq \delta^r / ((c_2 - c_3)\epsilon)^r \leq \delta^r / (c_3\epsilon)^r$ , and  $Q(B_i) - Q'(B_i^*) \leq \delta^r / (c_3\epsilon)^r$ , for any  $i = 1, \dots, k$ . As well,  $Q'(S^c) \leq \delta^r / (c_3\epsilon)^r$ . This implies that,

$$|Q(B_i^*)/Q(S) - Q'(B_i^*)/Q'(S)| = \left| Q(B_i) - \frac{Q'(B_i^*)}{1 - Q'(S^c)} \right| \leq \frac{2\delta^r / (c_3\epsilon)^r}{1 - \delta^r / (c_3\epsilon)^r} \leq 4\delta^r / (c_3\epsilon)^r,$$

where the last inequality holds as soon as  $\delta \leq c_3\epsilon/2^{1/r}$ . In short,  $W_r(Q, Q') \leq \delta$  implies that  $\|\Phi(Q) - \Phi(Q')\|_\infty \leq 4(\delta/c_3\epsilon)^r$ . By the same argument as that of Lemma 4.1

$$\begin{aligned}
\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) &= P_1(\{\mathbf{q} | \mathbf{q} \notin B_1; \|\mathbf{q} - \mathbf{q}'\|_\infty \leq 4(\delta/c_3\epsilon)^r \text{ for some } \mathbf{q}' \in B_1\}) \\
&\leq \frac{(1/2K)^{\alpha^* - 1} \Gamma(\alpha)}{\prod_{i=1}^K \Gamma(\alpha\beta_i)} (12(\delta/c_3\epsilon)^r)^{\alpha^*} / \alpha^* \times K(K-1)(6/\alpha^*)^{K-2} \\
&\leq \frac{2K^3 \Gamma(\alpha)}{\Gamma(\alpha^*)^K} \times 12^{\alpha^*} c_3^{-r\alpha^*} 6^{K-2} \alpha^{*1-K} (\delta/\epsilon)^{r\alpha^*}.
\end{aligned}$$

Here,  $\alpha^* = \min_{i \leq K} \alpha\beta_i = \min_{i \leq K} \alpha G(B_i) \leq \alpha \leq 1$ . Since  $\alpha^* \Gamma(\alpha^*) = \Gamma(\alpha^* + 1) \geq 1/2$ ,  $\Gamma(\alpha^*) \geq 1/(2\alpha^*)$ . Since  $\epsilon \geq c_1^{t+1} W_r(G, G')$  we obtain

$$\mathcal{D}(\mathcal{B}_\delta \setminus \mathcal{B}) \leq 24\Gamma(\alpha) (1/c_1^{t+1} c_3)^{r\alpha^*} K^3 12^K \alpha^* (\delta/W_r(G, G'))^{r\alpha^*}.$$

Finally, note that  $\alpha^* \leq \alpha/K$ . So,  $K^3 12^K \alpha^* \leq 24^K$ . Also,  $(1/c_1^{t+1} c_3)^{r\alpha^*} \leq \max\{1, (1/c_1^{t+1} c_3)^r\}$ , a constant dependent only on  $\mathcal{D}$ . In addition,  $K$  is non-increasing and  $g$  is non-decreasing, so  $K = K(\epsilon) \leq K(c_1^{t+1} W_r(G, G'))$ , and  $\alpha^* \geq \alpha g(c_1^{t+1} W_r(G, G'))$ . This allows further simplification to arrive at the desired bound.  $\square$

## 11 Proof of Lemma 7.4

The proof is organized in the following steps.

**Step 1** Let  $p_{Y|Q}$  denote the mixture density  $Q * f$ . Define the Hellinger information of  $W_2$  metric for a set  $\mathcal{Q} \subset \mathcal{P}(\Theta)$  given  $Q_0$ :

$$\psi_{\mathcal{Q}}(\delta) = \inf_{Q \in \mathcal{Q}; W_2(Q_0, Q) \geq \delta/2} h^2(p_{Y|Q_0}, p_{Y|Q}).$$

Also define  $\phi_{\mathcal{Q}} : \mathbb{R}_+ \rightarrow \mathbb{R}$  to be an arbitrary non-negative valued function such that for any  $\delta > 0$ ,  $\sup_{Q, Q' \in \mathcal{Q}; W_2(Q, Q') \leq \phi_{\mathcal{Q}}(\delta)} h^2(p_{Y|Q_0}, p_{Y|Q}) \leq \psi_{\mathcal{Q}}(\delta)$ . Define

$$B_K(Q_0, \delta) = \{Q \in \mathcal{P}(\Theta) | K(p_{Y|Q_0}, p_{Y|Q}) \leq \delta^2, K_2(p_{Y|Q_0}, p_{Y|Q}) \leq \delta^2\}. \quad (43)$$

The first step of the proof involves obtaining the following result: Suppose that there is a sequence  $\delta_n \rightarrow 0$  such that  $n\delta_n^2 \rightarrow \infty$ , a sequence of scalars  $M_n$ , a sequence of subsets of measures  $\mathcal{S}_n \subset \mathcal{P}(\Theta)$  and the following hold:

$$\begin{aligned} & \log D(\delta/2, \mathcal{S}_n \cap B_{W_2}(Q_0, 2\delta) \setminus B_{W_2}(Q_0, \delta), W_2) + \\ & \sup_{G_1 \in \mathcal{S}_n} \log D(\phi_{\mathcal{S}_n}(\delta), \mathcal{S}_n \cap B_{W_2}(G_1, \delta/2), W_2) \leq n\delta_n^2 \quad \forall \delta \geq \delta_n, \text{ a.s.} \end{aligned} \quad (44)$$

$$\frac{\Pi(\mathcal{P}(\Theta) \setminus \mathcal{S}_n | \mathcal{E}_n)}{\Pi(B_K(Q_0, \delta_n) | \mathcal{E}_n)} = o(\exp(-2n\delta_n^2)) \quad \text{a.s.}, \quad (45)$$

$$\frac{\Pi(\mathcal{S}_n \cap B_{W_2}(Q_0, 2j\delta_n) \setminus B_{W_2}(Q_0, j\delta_n) | \mathcal{E}_n)}{\Pi(B_K(Q_0, \delta_n) | \mathcal{E}_n)} \leq \exp[n\psi_{\mathcal{S}_n}(j\delta_n)/16] \quad (46)$$

for all  $j \geq M_n$ , a.s.

$$\exp(2n\delta_n^2) \sum_{j \geq M_n} \exp[-n\psi_{\mathcal{S}_n}(j\delta_n)/16] \rightarrow 0. \text{ a.s.} \quad (47)$$

Here, the almost sure statements are taken with respect to  $\Pi_G$ . Then, both Eq. (40) and (41) hold. The proof of this step is a straightforward modification of the proof of Theorem 4 in Nguyen [2013a] and is omitted.

**Step 2** By assumption (A2) and Lemma 3.2,  $h^2(p_{Y|Q}, p_{Y|Q'}) \leq K(p_{Y|Q}, p_{Y|Q'}) \leq C_1 W_r^r(Q, Q')$  for  $r \geq 1$ . Since

$$W_r^r(Q, Q') \leq \text{diam}(\Theta)^{r-1} W_1(Q, Q') \leq \text{diam}(\Theta)^{r-1} W_2(Q, Q'),$$

a valid choice for  $\phi_{\mathcal{S}}$  is  $\phi_{\mathcal{S}}(\delta) = \frac{\psi_{\mathcal{S}}(\delta)}{4C_1 \text{diam}(\Theta)^{r-1}}$ . Recall the following facts (Proposition 1 of Nguyen [2013a]): If  $f$  is an ordinary smooth density function on  $\mathbb{R}^d$  with parameter  $\beta > 0$ , for any  $d' > d$ ,  $\psi_{\mathcal{P}(\Theta)}(\delta) \geq c(d, \beta) \delta^{4+2\beta d'}$  for some constant  $c(d, \beta) > 0$ . If  $f$  is a supersmooth density function on  $\mathbb{R}^d$  with parameter  $\beta > 0$ ,  $\psi_{\mathcal{P}(\Theta)}(\delta) \geq \exp[-c(d, \beta) \delta^{-\beta}]$  for some constant  $c(d, \beta) > 0$ . These give immediate lower bounds on  $\phi_{\mathcal{S}}(\delta) \geq \phi_{\mathcal{P}(\Theta)}(\delta)$ . Note in the following that constants  $c$  may be different in each appearance.

**Step 3** Consider the ordinary smooth case. We shall construct a sequence of subsets  $\mathcal{S}_n$  and a sequence  $\delta_n \rightarrow 0$  such that both Eq. (44) and Eq. (45) hold. In fact, we set  $\mathcal{S}_n := \mathcal{B}_n$ , whose existence and properties are given by Lemma 7.3. Note that the choice of  $\mathcal{B}_n$  is random because it depends on  $G$ , in addition to  $\delta_n$ . For any  $\delta \geq \delta_n$ ,  $\Pi_G$ -almost surely we have

$$\begin{aligned} & \log D(\delta/2, \mathcal{S}_n \cap B_{W_2}(Q_0, 2\delta) \setminus B_{W_2}(Q_0, \delta), W_2) \\ & \leq \log N(\delta/4, \mathcal{B}_n, W_2) \\ & \leq k[(\log N(\delta_n/16, \Theta, \|\cdot\|) + \log(e + 16e \text{diam}(\Theta)^2/\delta_n^2))] \\ & \lesssim kd \log(\text{diam}(\Theta)/\delta_n). \end{aligned} \quad (48)$$

In addition,  $\Pi_G$ -almost surely, for any  $\delta \geq \delta_n$ ,

$$\begin{aligned} & \sup_{G_1 \in \mathcal{S}_n} \log D(\phi_{\mathcal{S}_n}(\delta), \mathcal{S}_n \cap B_{W_2}(G_1, \delta/2), W_2) \\ & \leq \log D(\phi_{\mathcal{S}_n}(\delta), \mathcal{S}_n, W_2) \leq \log N(c\delta^{4+2\beta d'}, \mathcal{B}_n, W_2) \\ & \leq k[\log N(c\delta_n^{4+2\beta d'}/4, \Theta, \|\cdot\|) \\ & \quad + \log(e + 4e \text{diam}(\Theta)^2/c\delta_n^{8+4\beta d'})] \\ & \lesssim kd^2 \log(\text{diam}(\Theta)/\delta_n). \end{aligned} \quad (49)$$

From Eqs. (48) and (49), by setting  $\delta_n = M_0(\log n/n)^{1/2}$ , where  $M_0$  is a positive constant depending on  $k, d, \text{diam}(\Theta)$  and  $\beta$ , the entropy condition expressed by Eq. (44) immediately follows.

According to part (b) of Lemma 7.3,  $\Pi(\mathcal{P}(\Theta) \setminus \mathcal{S}_n | \mathcal{E}_n) \leq \epsilon_n \text{diam}(\Theta)/\delta_n^2$ . By Lemma 7.1, if  $\delta_n$  satisfies  $\delta_n / \log(1/\delta_n) \gtrsim \epsilon_n^{r/2}$ , then  $\Pi_G$ -almost surely,

$$\Pi(B_K(Q_0, \delta_n) | \mathcal{E}_n) \geq c(\delta_n / \log(1/\delta_n))^{2(\alpha+k-1)}.$$

It follows that if  $\delta_n \gtrsim \epsilon_n^{r/2} \log(1/\epsilon_n)$ , then

$$\frac{\Pi(\mathcal{P}(\Theta) \setminus \mathcal{S}_n | \mathcal{E}_n)}{\Pi(B_K(Q_0, \delta_n) | \mathcal{E}_n)} \leq \epsilon_n \delta_n^{-2(\alpha+k)} [\log(1/\delta_n)]^{2(\alpha+k-1)} \text{diam}(\Theta)/c. \quad (50)$$

From Eq. (50), the condition expressed by Eq. (45) is verified if

$$\begin{aligned} \delta_n & \gtrsim \epsilon_n^{r/2} \log(1/\epsilon_n) \quad \text{and} \\ \delta_n^{2(\alpha+k)} & \gtrsim \epsilon_n \exp(4n\delta_n^2) (\log(1/\delta_n))^{2(\alpha+k-1)}. \end{aligned}$$

It is easily seen that both inequalities hold for the rate  $\delta_n = M_0(\log n/n)^{1/2}$ , if  $\epsilon_n \lesssim \min\{(n \log n)^{-1/r}, n^{-(\alpha+k+4M_0)} (\log n)^{-(\alpha+k-2)}\}$ . In other words, if  $\epsilon_n$  tends to 0 sufficiently fast, we obtain parametric rate of convergence  $(\log n/n)^{1/2}$  for the posterior of  $Q * f$ .

It remains to verify Eqs (46) and (47). It suffices to construct a sequence of  $M_n$  so that  $\Psi_{\mathcal{S}_n}(M_n \delta_n) \geq C \delta_n^2$  for a large constant  $C > 0$ . Since  $\Psi_{\mathcal{S}_n}(\delta) \gtrsim \delta^{4+2\beta d'}$  for any  $d' > d$ ,  $M_n$  can be chosen so that  $M_n \delta_n \gtrsim \delta_n^{\frac{1}{2+\beta d'}}$ .

**Step 4** We turn to the case of supersmooth kernel density  $f$ .  $\mathcal{S}_n = \mathcal{B}_n$  is constructed in the same way as in Step 3. Bounds (48) and (50) remain valid, but Eq. (49) is replaced (using the bound on  $\phi_{\mathcal{S}_n}$  described in Step 2): for any  $\delta \geq \delta_n$ ,

$$\begin{aligned}
& \sup_{G_1 \in \mathcal{S}_n} \log D(\phi_{\mathcal{S}_n}(\delta), \mathcal{S}_n \cap B_{W_2}(G_1, \delta/2), W_2) \\
& \leq \log D(\phi_{\mathcal{S}_n}(\delta), \mathcal{S}_n, W_2) \leq \log N(\exp[-c\delta^{-\beta}], \mathcal{B}_n, W_2) \\
& \leq k[\log N(\exp[-c\delta_n^{-\beta}]/4, \Theta, \|\cdot\|) \\
& \quad + \log(e + 4e \operatorname{diam}(\Theta)^2 \exp[c\delta_n^{-\beta}])] \lesssim kd\delta_n^{-\beta}.
\end{aligned} \tag{51}$$

From Eqs. (48) and (51), by setting  $\delta_n = M_0(1/n)^{1/(\beta+2)}$ , where  $M_0$  is a large constant depending on  $k, d, \operatorname{diam}(\Theta)$ , the entropy condition expressed by Eq. (44) immediately follows. As in the previous step, if  $\epsilon_n$  tends to 0 sufficiently fast, i.e.,  $\epsilon_n \lesssim n^{-2(\alpha+k)/(\beta+2)}(\log n)^{-2(\alpha+k-1)} \exp(-4n^{\beta/(\beta+2)})$ , then Eq. (45) is satisfied for the given choice of  $\delta_n$ . That is, we obtain a parametric rate  $(1/n)^{1/(\beta+2)}$  of posterior concentration of  $Q * f$  that does not depend on  $d$ .

**Step 5** The previous two steps established (iv) and (v), i.e., deriving rates  $\delta_n$  when  $\epsilon_n$  tends to 0 sufficiently fast. In the final step we derive  $\delta_n$  that is valid for any  $\epsilon_n$ . Here, we turn to the choice  $\mathcal{S}_n := \mathcal{P}(\Theta)$ : this is in fact a convex set, so we can appeal to Theorem 3 of Nguyen [2013a], which avoids having to upper bound the covering number in terms of  $W_2$  radius  $\phi_{\mathcal{S}_n}(\delta)$ . (Indeed, a similar calculation has been carried out in their Theorem 6, which gives posterior concentration rates of mixing measures for the stand-alone mixture model  $Q * f$ , which yields the rate  $(\log n/n)^{1/(d+2)}$ ). We omit the similar derivations here, which give  $\delta_n \asymp (\log n/n)^{1/(d+2)} + \epsilon_n^{r/2} \log(1/\epsilon_n)$  as the posterior concentration rate of the mixture density  $Q * f$ . The extra quantity depending on  $\epsilon_n$  is needed so that Lemma 7.1 can be applied.

## References

- A. Barron, M. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27:536–561, 1999.
- J. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition, 1993.
- D. Blackwell and J. MacQueen. Ferguson distributions via polya urn schemes. *Annals of Statistics*, 1:353–355, 1973.
- P. Buhlmann and S. van de Geer. *Statistics for high-dimensional data: Methods, theory and applications*. Springer, 2011.
- R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of American Statistical Association*, 83:1184–1186, 1988.

- H. Doss and T. Sellke. The tails of probabilities chosen from a Dirichlet prior. *Annals of Statistics*, 10(4):1302–1305, 1982.
- K. J. Falconer. *The geometry of fractal sets*. Cambridge University Press, 1985.
- J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272, 1991.
- T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1: 209–230, 1973.
- I. Garcia, U. Molter, and R. Scotto. Dimension functions of Cantor sets. *Proceedings of the American Mathematical Society*, 135(10):3151–3161, 2007.
- S. Ghosal. Dirichlet process, related priors and posterior asymptotics. In N. Hjort, C. Holmes, P. Mueller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, Cambridge, UK, 2010.
- S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.
- S. Ghosal, J. K. Ghosh, and A. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.
- J. K. Ghosh and R. V. Ramamoorthi. *Bayesian nonparametrics*. Springer, 2002.
- E. Giné and R. Nickl. Rates of contraction for posterior distributions in  $l^r$ -metrics,  $1 \leq r \leq \infty$ . *Annals of Statistics*, 39:2883–2911, 2011.
- N. Hjort, C. Holmes, P. Mueller, and S. Walker (Eds.). *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010.
- R. Korwar and M. Hollander. Contributions to the theory of Dirichlet processes. *Annals of Probability*, 1(4):705–711, 1973.
- E. Lehmann and G. Casella. *Theory of point estimation*. Springer, 1998.
- X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 41(1):370–400, 2013a.
- X. Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, invited revision, 2013b.
- J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B*, 73(5):689–710, 2011.
- J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

- X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29:687–714, 2001.
- Y. W. Teh and M. I. Jordan. Hierarchical bayesian nonparametric models with applications. In N. Hjort, C. Holmes, P. Mueller, and S. Walker, editors, *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, Cambridge, UK, 2010.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.
- A. van der Vaart and J. van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *Annals of Statistics*, 36(3):1435–1463, 2008.
- A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.
- Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.
- S. Walker. New approaches to bayesian consistency. *Ann. Statist.*, 32(5):2028–2043, 2004.
- S. Walker, A. Lijoi, and I. Prunster. On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.*, 35(2):738–746, 2007.
- W. H. Wong and X. Shen. Probability inequalities for likelihood ratios and convergences of sieves mles. *Ann. Statist.*, 23:339–362, 1995.
- C. Zhang. Fourier methods for estimating mixing densities and distributions. *Annals of Statistics*, 18(2):806–831, 1990.